Research Ideas and Outcomes 1: e7573
doi: 10.3897/rio.1.e7573

OPEN ACCESS

H2020 Grant Proposal

Enabling Open Science: Wikidata for Research (Wiki4R)

Daniel Mietchen[‡], Gregor Hagedorn[‡], Egon Willighagen[§], Mariano Rico[|], Asunción Gómez-Pérez[|], Eduard Aibar[¶], Karima Rafes[#], Cécile Germain[¤], Alastair Dunning[«], Lydia Pintscher[»], Daniel Kinzler[»]

‡ Museum für Naturkunde Berlin, Berlin, Germany
§ Maastricht University, Maastricht, Netherlands
| Universidad Politécnica de Madrid (UPM), Madrid, Spain
¶ Universitat Oberta de Catalunya, Barcelona, Spain
# Inria, Paris, France
¤ Université Paris-Sud, Paris, France
« The European Library, Den Haag, Netherlands
» Wikimedia Deutschland e.V., Berlin, Germany

Corresponding author: Daniel Mietchen (daniel.mietchen@mfn-berlin.de)

Reviewable v1

Received: 21 Dec 2015 | Published: 22 Dec 2015

Citation: Mietchen D, Hagedorn G, Willighagen E, Rico M, Gómez-Pérez A, Aibar E, Rafes K, Germain C, Dunning A, Pintscher L, Kinzler D (2015) Enabling Open Science: Wikidata for Research (Wiki4R). Research Ideas and Outcomes 1: e7573. doi: 10.3897/rio.1.e7573

Abstract

Wiki4R will create an innovative virtual research environment (VRE) for Open Science at scale, engaging both professional researchers and citizen data scientists in new and potentially transformative forms of collaboration. It is based on the realizations that (1) the structured parts of the Web itself can be regarded as a VRE, (2) such environments depend on communities, (3) closed environments are limited in their capacity to nurture thriving communities. Wiki4R will therefore integrate Wikidata, the multilingual semantic backbone behind Wikipedia, into existing research processes to enable transdisciplinary research and reduce fragmentation of research in and outside Europe. By establishing a central shared information node, research data can be linked and annotated into knowledge. Despite occasional uses of Wikipedia or Wikidata in research, significant barriers to broader adoption in the sciences or digital humanities exist, including lack of integration into existing research processes and inadequate handling of provenances. The proposed actions include providing best practices and tools for semantic mapping, adoption of citation and author identifiers, interoperability layers for integration with existing

research environments, and the development of policies for information quality and interchange. The effectiveness of the actions will be tested in pilot use cases. Unforeseen barriers will be investigated and documented. We will promote the adoption of Wiki4R by making it easy to use and integrate, demonstrate the applicability in selected research domains, and provide diverse training opportunities. Wiki4R leverages the expertise gained in Europe through the Wikidata and DBpedia projects to further strengthen the established virtual community of 14000 people. As a result of increased interaction between professional science and citizens, it will provide an improved basis for Responsible Research and Innovation and Open Science in the European Research Area.

Keywords

Virtual Research Environment, Wikidata, identifiers, citizen science, collaboration, concept mapping, ontologies

Context

This article describes a research proposal that was submitted on January 14, 2015, to the European Commission's H2020-EINFRA-2015-1 call for proposals for e-Infrastructures for virtual research environments (VRE). Its main proposition was to build this VRE by integrating research workflows with the Web through Wikidata, an existing open environment for managing structured information collaboratively across all domains of human knowledge.

Wikidata is being built entirely using open software and open content, and it had a community of over 14,000 monthly contributors worldwide at the time, which has since risen to over 16,000. A VRE built on that basis would thus avoid some of the problems traditionally plaguing many VREs: software and/ or content that are proprietary and hence siloed, and a lack of community uptake.

Moreover, Wikidata's public version histories of both content and software (another feature missing in many VREs) provide for a solid basis for open science, which emerged as a new priority of the European Commission's research activities (and elsewhere) over the course of the year.

The project proposed to prototype the integration of Wikidata with a set of workflows that are either transdisciplinary (e.g. for handling scholarly references) or particular to specific use cases like chemistry or mineralogy. Apart from project management (WP 1), the individual workpackages then focused on

- (WP 2) describing research-relevant concepts (e.g. molecules, chemical reactions or journal articles) in terms of Wikidata's continuously evolving data model and with a view to maximize compliance with Web standards;

- (WP 3) establishing exchange and curation workflows between Wikidata and external databases hosting information about these concepts (e.g. identifier mapping);
- (WP 4) using Wikidata content in research workflows (e.g. querying it, or using its identifiers in research notebooks), including citizen science projects;
- (WP 5) training, education and outreach around these topics, with a focus on creating openly licensed educational materials that could be reused in other contexts.

Several of the activities proposed in the framework of the project have since experienced progress due to support from different parts of the community. For instance, there are now several mechanisms to query Wikidata in real-time or nearly so, some of which are using SPARQL (cf. Task 4.1). This allows, for instance, to get a list of countries ordered by the number of their cities with a female mayor. Basic support for units is now available too (cf. Task 2.1), and the integration of bibliographic metadata of scholarly citations (part of Task 3.1) is moving forward.

On the other hand, some of the proposed activities have seen little to no progress over the year, although we continue to see them as valuable for the community. These include, for instance, the integration of the European Library's dataset of bibliographic metadata for books into Wikidata (another part of Task 3.1) or Task 3.3, concerned with

- identifying potential external data sources suitable for integration with Wikidata and
- exploring the benefits for data providers to engage in such an open sharing of their data.

or Task 4.4, which was about

- involving Wikidata in scientific curation workflows
- connecting Wikidata with citizen science projects

as well as the reuse of data from the Horizon 2020 Open Data Pilot (part of Task 3.1), or most of WP 5, e.g. the development of online tutorials and course materials.

Some training events (cf. Task 5.5) have already been organized by the community, e.g. at the Semantic Web applications and tools for life sciences (SWAT4LS) conference earlier this month by the Gene Wiki team, whose NIH-supported work on Wikidata had inspired our Wiki4R proposal and who also received that conference's Best paper prize for their contribution "Wikidata: A platform for data integration and dissemination for the life sciences and beyond".

Like most of the submissions to this heavily oversubscribed call, our proposal was rejected. While this may have been appropriate in the context of this specific call (hard to tell, since there is no public information about which projects were submitted, and why the winning ones were chosen over the others), we think that the ideas we put into the proposal are still worth pursuing, and we want to encourage others to build on our proposal to move forward with the integration of research workflows with Wikidata.

There are some signs that this may indeed be happening: the WikiProject Wikidata for Research, which was initially started about a year ago in order to facilitate the open and collaborative drafting of our proposal (including assembling the group of project partners) has developed into a platform for sharing information at the interface between Wikidata and research, and about one third of its ca. 50 participants have signed up after our proposal was submitted. Some of them have, independent of us, explored how medical content on Wikipedia could benefit from closer integration with scholarly databases through Wikidata (similar to Task 3.1), and published a paper about it, while the Wikimedia Foundation recently funded a proposed project to mine selected external websites for facts that could be turned into statements on Wikidata, with the URL of that external site serving as a reference (similar to Task 3.2). Of note, a number of external websites and services have started to use Wikidata identifiers in their workflows (cf. Task 4.3), many of them in the area of cultural heritage (cf. Task 4.5).

In preparing the proposal for publication, we kept as closely as possible to the original text (which is still available as a preprint on Zenodo) and whose DOI was included in the original submission. Bothe the preprint and this version are missing the letters of support, since the logos in the letterheads are under copyright of the respective institutions and cannot be posted under an open license. Similarly, the reviews cannot be posted here, since we do not know who has the copyright, and whether they would consent to this publication. We encourage funders to make it simpler for applicants to share the reviews of their proposals, and we will update this article should we receive permission to post the reviews.

Except for adding this Context section, the changes made here were the addition of some metadata sections (e.g. Keywords, Funding program) and a few copyedits (typos and making sure all tables and figures are actually cited in the text). We also added an Acknowledgements section and dropped Anonymous as co-author in order to comply with journal policy.

Excellence

Participants are listed in Table 1.

Table 1.
List of participants

Participant No	Participant organisation name	Country
1 (Coordinator)	Museum für Naturkunde Berlin (MfN)	Germany
2	Universidad Politécnica de Madrid (UPM)	Spain
3	Maastricht University (UM)	Netherlands
4	Wikimedia Deutschland (WMDE)	Germany
5	Universitat Oberta de Catalunya (UOC)	Spain

6	Europeana Foundation (EF)	Netherlands
7	Université Paris Sud (UPS)	France

Objectives

The overarching goals of the proposed project are:

1. To support the application of Open Science principles[1] by leveraging and strengthening an existing trusted, shared, long-term-sustainable, collaborative platform for the curation of open data.
2. To enable researchers across Europe and beyond to perform transdisciplinary research, overcoming the fragmentation of virtual research environments (VREs) and separated data silos by moving to an open, shared data environment with collaborative data curation.
3. To support increasingly rich interlinking of information and digital knowledge representations that can be used by both humans and machines.
4. To increase the capabilities of both professional and citizen scientists and the capacities of their organisations to collaborate with each other in mutually beneficial and potentially transformative ways.
5. To support the development of Open Science policies and, through increased dialogue between scientists and citizens, foster Responsible Research and Innovation in the European Research Area.

[1]We will use the terms Open Science and Open Research interchangeably here, with the understanding that no domain of research – academic, industrial, governmental or other – should be regarded as excluded from using an open approach. We will underline that by running this transdisciplinary project itself as an open project, where not only the results but also the process of research are made transparent if feasible.

To achieve these goals, we will build a VRE on the basis of Wikidata, the database that anyone can edit, which serves as the semantic backbone for structured data in Wikipedia and its sister projects. It represents a massive development step and paves the way towards a structured future of Wikipedia. Upon its launch, Sue Gardner, Executive Director of the Wikimedia Foundation, wrote: “Before Wikidata, Wikipedians needed to manually update hundreds of Wikipedia language versions every time a famous person died or a country’s leader changed. With Wikidata, such new information, entered once, can automatically appear across all Wikipedia language versions. That makes life easier for editors and makes it easier for Wikipedia to stay current."

Wikidata is a major open data platform for massive online collaboration and is designed to share “the sum of all human knowledge”, including all domains of scholarly and scientific knowledge. It has already become a major focus point for sharing scholarly as well as technical information. As a key element of the wider landscape of citizen-driven open knowledge initiatives, Wikidata is unfolding with the active participation of a global and

multilingual community of volunteers – more than 14,000 of them contribute to the project on a monthly basis.

By building on Wikidata, the proposal avoids the problems of many other VRE platforms that struggle to find sufficiently large communities of users. An ecosystem of infrastructure, technologies and tools already exists, but the interactions with professional research are insufficiently developed. Most professional science and citizen science projects presently work on completely separate digital platforms. The present project addresses known barriers and investigates the yet unknown ones. By adding innovative features required for research workflows, providing help with semantic mapping, providing best practices and training materials, and researching pilot use cases, this project will enable the Wikidata platform to develop from a virtual environment to a virtual research environment for Open Science, engaging both professional researchers and citizen data scientists in new and potentially transformative forms of collaboration.

The specific objectives of the project are:

1. Goal 1:

1.1. to provide a VRE where contributions are transparent and trusted because all changes are accountable and a full history of changes is publicly available;

1.2. to enhance the verifiability of statements by accompanying them with provenance and references (Task 3.2);

1.3. to increase the scale and immediacy with which scientific information is made globally available to both people and machines, and curated by them;

2. Goal 2:

2.1. to support transdisciplinary science by promoting the adoption of a single open data VRE platform across disciplines from the natural sciences to the humanities;

2.2. to establish a single open-data VRE platform by piloting it for a set of core use cases (chemistry, library and information science, biodiversity science, mineralogy and the cultural heritage sector);

2.3. to promote the open-source Wikibase software (of which Wikidata is an installation) as a framework also installable and adaptable for institution-internal knowledge management;

3. Goal 3:

3.1. to establish the use of stable identifiers for scientific objects and concepts (which are dereferenceable, i.e. can be actively followed to obtain further information) in a pilot of 5 large and 10 smaller scientific datasets (WP2 and Task 3.1);

3.2. to increase the interlinking of information by making a large number of scientific objects and concepts citable in a long-term stable and sustainable manner in the Linked Open Data Cloud;

4. Goal 4:

4.1. to develop and provide documentation, training materials, tutorials, open educational course materials, best-practices advice, as well as dissemination and community engagement events (WP5);

4.2. to increase the number of citizen data scientists and citizen data curators interacting with professional scientific researchers and professional research organisations with the aim to increase the quality of the data available for all (Task 4.5);

4.3. to increase the number of external citizen science projects that use Wikidata (Task 4.5);

5. Goal 5:

5.1. to participate in the EC's Open Data Pilot as both provider and re-user of data from multiple domains, which can inform quality improvement measures for the Pilot;

5.2. to analyse the motivations for the open sharing of data in research contexts, identify best practices, and distill recommendations on the design and implementation of institutional and overarching data-sharing policies.

Relation to the work programme

The proposal is strongly aligned with the specific challenges of the Horizon 2020 call for "e-Infrastructures for virtual research environments (VRE)". It will address the challenge of capacity-building in interdisciplinary European research by adapting a trans-disciplinary collaborative open-data knowledge platform to the needs of professional researchers (WP2–4), developing best practices for and providing training around that (WP5). Being web-based and supporting standard forms of knowledge expression, knowledge integration and computing, the proposed VRE will integrate resources across all layers of the e-infrastructure. Its strengths will be full and inherent transdisciplinarity and mechanisms to support openness with respect to standards, transparency and accountability; options for dissent; and mechanisms to encourage consensus-building. It will provide significant computing resources to improve the analysis and integration options of both existing general knowledge and the newly integrated scientific data.

The Wiki4R VRE will be modular. The existing Wikidata infrastructure already fulfils several requirements of a modern VRE:

- abstraction from underlying infrastructures,
- the use of open source software with a large global developer base,
- the use of globally accessible, well-documented interfaces and APIs,

- the use of web-based workflows,
- a service-oriented architecture supporting both human and machine use,
- an ecosystem of tools interacting with the core infrastructure, and
- full interoperability with standard Semantic Web technologies.

The action will specifically address the presently tentative connection between Wikidata stakeholders (open-source volunteer developers from civil society) and professional researchers and their organisations. With its limited resources, the project will target on-platform development that is beneficial for research across domains. However, we expect that future extensions – by both the project partners and others – will create an ecosystem of domain-specific functionality around this VRE through both on- and off-platform enhancements. Wikidata and DBpedia will provide generic data storage, integration, curation and analysis services to a multilingual open-data user community, with strong attention to trust, provenance, accountability, and verifiability. All data on Wikidata will be open-access. Domain- and discipline-specific services from public, private and commercial research institutions will build on this foundation.

The Wiki4R VRE will be relevant for data-oriented research that addresses a broad range of societal challenges, from the natural sciences (including health, climate research, environment, agriculture and forestry), to engineering (including energy and transport), mathematics, information science, the digital humanities, education and unsupervised learning. The project partners span major branches of the natural sciences, the arts and humanities, the cultural and natural heritage sector, and civil society, along with partners with information science and Semantic Web expertise. Due to the open and participatory nature of Wikidata, people and organisations outside the partner consortium will be free to contribute. Challenges regarded as important by citizens will be addressable: no restrictions will be imposed by the consortium partners on the type and usage of data.

The greatest strength of the action will be new ways of a wide spectrum of citizens in research, data analysis and knowledge sharing, leading to a more inclusive, innovative and reflective European society. Open knowledge and the engagement of citizens and society in a responsible research and innovation process resonate strongly with European values. Global initiatives with a European basis – like the Open Knowledge Foundation, Wikimedia's Wikidata and DBpedia – provide very valuable contributions to that.

Concept and approach

The research agenda for VREs by Candela et al. (2013) highlights that usability, sustainability and the reuse of services and resources should be built into the design, and that VREs should be integrated with other existing infrastructure in a mutually beneficial way, enhancing their sustainability and their value in the eyes of broader research communities. Promoting a VRE user community early in development will be critical to ensuring sufficient uptake for the sustainability of VREs and the potential for future research on improving them. Ideally, the VRE should come to be seen as an essential technology for the target community; this will require both social and technical innovation. As Candela et al. suggest, the "focus should be primarily on using technology to identify

and rationalise workflows, procedures, and processes characterising a certain research scenario rather than having technology invading the research scenario and distracting effort from its real needs."

Many existing VREs are designed as secure and closed "remote desktop" systems, in which a selected and managed number of researchers can work, and which provide feature-complete, controlled and customised data access and computing services to the larger world. This approach clearly has its applications; but it typically does not scale well to a large number of researchers with diverse research needs. One reason for this concerns the management of access and collaboration rights. If researchers fall into few well-defined groups with standard tasks (e.g. citizen scientists with homogeneous data collection rights), large numbers can be managed. However, researchers typically have very diverse needs. This can extend to citizen scientists, if they are considered as partners who collaborate in the full research process. The rights management for a VRE with a large number of users – possibly 100,000 – could become prohibitively expensive. Another reason concerns VRE-specific methods of access to external data and services, the sum of which can be expensive to implement and sustain. This limits this type of VREs with respect to the number and diversity of research questions and typically results in discipline-specific VREs. Finally, while the results from such a VRE may be channelled into open-access publications, such VREs typically do not easily lend themselves to open science, where the research process itself is shared (Fig. 1).

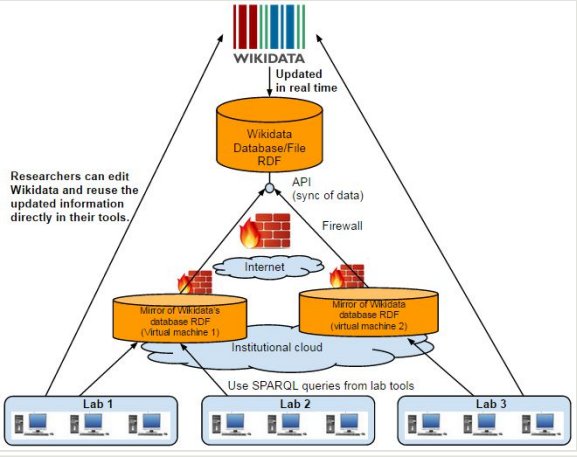

Figure 1.

Outline of envisioned platform where Wikidata content can be used within an institutional firewall.

Our proposal does not attempt to build such a secured, highly customised, discipline-specific research environment with expensive-to-maintain, purpose-tailored tools. Instead, it focuses on the needs of open science and empowering researchers to work together across disciplines in an open environment. Ultimately, the web itself – especially the Semantic Web combined with web-based computing services – is the full VRE. Specific

components, such as the Wikidata–science connection developed here, are components of this VRE. They provide services to the web, but do not limit the use of the web and its services.

The concept of open science (alternatively called “open research” if disciplines like philosophy, history, are not included in the definition of “science”) is central to this proposal. Open science is highly inclusive, inviting collaboration from professional peers as well as other interested parties, including citizen scientists. It is also open with respect to the process, inviting collaboration and providing access to research data as soon as they are created.

Managing openness is not a trivial exercise. It requires both appropriate technical foundations and procedural and social competencies, rules and policies. Openness and collaboration need to be maximised, dissent needs to be possible, consensus needs to be encouraged, all actions need to be owned, and all actors held accountable. Contributions must be not only technically documented, but also suitably acknowledged with provenance-related information that may be available to both humans and machines, yet transparent to the search process for content.

Many of the technologies and policies have already been developed in the context of Wikipedia and Wikidata. However, there is no automatic adoption of these techniques by professional science. We identify social barriers, especially a lack of examples and lack of training, but also technical gaps. The project therefore studies these problems. It works both on enhanced technical solutions and on providing professional researchers with examples, guidelines and best practice documentation as to how their data can be shared openly and integrated with other open data (see “Optimizing openness”, **T3.3**).

The Wiki4R VRE, which will be tested in a number of pilot cases, supports web-based, discipline-neutral, standard analysis methods. It empowers researchers to share data widely, to integrate and interlink research data with general, non-discipline-specific data, to cite data and have their own data cited. Finally, it empowers researchers to share the burden of data curation between professional science organisations and citizen data scientists/citizen data curators.

The open-science approach will not work for all possible domains of systematic inquiry (e.g. industrial research aiming at patentable results). If legal, data privacy, commercial, or other considerations prevent openness, it will be necessary to design mechanisms that allow certain data to be stored privately. Aspects of data privacy will be carefully considered where applicable, and the Wikimedia Foundation’s privacy policy will be enforced.

However, where open science is possible, it will be highly desirable. By increasing opportunities for collaboration, by increasing transparency and trust, by providing research progress and feedback opportunities as early as possible, and by providing more examples for students to learn from, research and education will become more efficient.

Factors that facilitate the uptake of the Wiki4R VRE for open science use are open data, coverage of all domains of knowledge, a powerful API and strong support for multilingualism (in terms of the data, the community, and the user interface). The use of a CC0 license will avoid forced attribution stacking for primary data, while at the same time mechanisms support and encourage proper scholarly citation. The deep integration of Wikidata with Wikipedia and its sister projects provides an excellent basis for recognition, long-term sustainability, and high impact. The consortium expects to have fewer issues with the credibility, sustainability and longevity concerns most EU projects typically fight with. One limitation that the present action will address is a lack of acceptance as a VRE for professional research. A key goal of the actions will be to ensure that the Wiki4R VRE will be readily accepted by many partners outside the consortium. The number of researchers targeted by this action is therefore potentially very high.

An important action is the integration of semantic entities and ontologies between research data and existing Wikidata items and properties to improve interoperability (WP2). The study and alignment of ontologies in this context is based on the realisation that research is dynamic: in the foreseeable future, no single ontology will be able to cover the needs of all diverse research groups. Even within disciplines, ontologies will be constantly evolving. Wikidata therefore supports an agile development of multiple ontologies, including the option that ontologies developed by different researchers may be contradictory. The choice of a single ontology will often be required for analysis purposes. The fundamental system is not built for a priori “truth”, but for discourse and research. Thus, instance/subclass assertions will underlie the same requirements for ownership and provenance reference as all other statements.

Sharing the semantics of research data early in the research process has great potential in increasing the efficiency of research. Doing so enables the expression of knowledge by the authoritative researchers, rather than the kind of guesswork that typically has to occur when adding semantic tagging to research publications that are not inherently semantic (but, for example, have been published as PDF). Another problem addressed is the problem that identifiers for concepts or objects may be required for further work long before the formal results are finally published. Even without full open science, much is being reported on conferences and in communications that influences the research of other groups and makes it desirable to semantically refer to new objects and concepts as early as possible. The creation and dissemination of appropriate identifiers should be one of the first steps in an open-science process (Fig. 2).

The proposed social and technical design of the Wiki4R VRE addresses the needs and requirements of a diverse network of project partners and associate partners, who offer use cases for data deposition and reuse as well as provision of value-added services. There are still comparatively few examples of Wikidata use in research contexts, therefore the project will include the development of new projects as well as associated tools and infrastructure that support these and existing use cases. Several partners are specifically interested in the integration of heterogeneous data from multiple sources, for example:

Figure 2.

Prototype of the platform at the Center for Data Science of Paris-Saclay, where Wikidata identifiers are already used to link public data and public research data.

- providing bibliographic metadata across scholarly domains for direct linking with research data
- integration of data across chemistry, between chemistry and biology and between chemistry and mineralogy
- integration of metadata about heritage collections
- integration of multilingual data.

The Wiki4R VRE will exist in an ecosystem of tools, services and infrastructures at national and European levels. Several parts of the ecosystem have been developed by project partners, maximising the ease with which the present project can build upon these developments. Examples include The European Library (part of the Europeana Foundation , providing access to the catalogues of 48 National and Research Libraries of Europe) and Open PHACTS (providing pharmacological data to aid drug discovery). Further examples are portals for advanced data exploration, multilingual integration, and semantic linking of data. Examples of transdisciplinary and multilingual data-driven research include analysing relationships between data items, such as the social and intellectual relationships between ancient philosophers, linkages between scientific facts extracted from the scholarly literature, or the detection of biases across many languages (see also gender bias examples further down). Value-added data mining services such as ContentMine have already defined use cases for automated deposition and retrieval of data in the context of a social architecture. They combine use of structured data from Wikidata to find connections between incoming data streams from the scientific literature with community curation. A number of visualisation projects use Wikidata, for example to analyse networks of ancient philosophers.

The existing Wikidata community includes domain specialists such as the editors of WikiProject Chemistry. In many cases, this means that mapping and external links to relevant resources already exist, of which the project can take advantage.

The following research innovation activities are especially relevant for Wiki4R:

The **Europeana Foundation** and **The European Library** (already mentioned above) are supported as part of the EU's Digital Agenda and expected to be funded as part of the EU's Connected Europe Facility. The Europeana Foundation acts as the central aggregator for data about Europe's culture heritage. Europeana has aggregated 32 million records related to digitised objects in Europe's museums, archives and libraries. For this purpose, Europeana has developed a Europeana Licencing Framework to harmonise the rights statements about digitised content and a related metadata model (the Europeana Data Model) for creating interoperable metadata from different domains. These two pillars ensure that the data available from Europeana - which will put onto Wikidata as part of this project - is all marked as CC0 and available as standardised Linked Open Data.

In addition to the central aggregator, various other specific aggregators focus on specific areas of the cultural sector (e.g. museums or archives). In the context of this proposal, The European Library will make its dataset of over 90 million bibliographic records available, i.e. data about published books. The data is drawn from national and research libraries across Europe - in essence, the European Library dataset acts as the record of publication in Europe. Similar to Europeana, the dataset is available as CC0 and as standardised LOD of national, regional, and local endeavours where complementarities and new challenges are clearly identified and acted upon.

RENDER is a finished FP7 project which provided a comprehensive conceptual framework and technological infrastructure for enabling, supporting, managing and exploiting information diversity in Web-based environments. Diversity was viewed as a crucial source of innovation and adaptability, ensuring the availability of alternative approaches towards solving hard problems, and provides new perspectives and insights on known situations. Equally important, embracing diversity in information management is essential for enhancing state-of-the-art technology in this field with novel paradigms, models, and methods and techniques for searching, selecting, ranking, aggregating, clustering and presenting information purposefully to users, thus alleviating critical aspects of information overload. The concepts, methods, techniques and technology developed by RENDER will especially inform the diversity approach for Wiki4R.

Gene Wiki (an Associate partner, via Andrew Su at Scripps) is a highly relevant project for our purpose in that GeneWiki successfully implements the dissemination of professional research information into Wikipedia and increasingly Wikidata. Its focus is on human genes, diseases and drugs, which has some very limited thematic overlap with our activities on metabolites (see also **T3.1**). Methodologically, we envisage a close collaboration, e.g. in terms of sharing in the development of code for curating Wikidata items and in collaborating with relevant WikiProjects in the semantic modelling parts **(WP2).** There is also a conceptual overlap in that some of our proposed activities map well

onto activities that were contained in the Gene Wiki proposal but not approved for funding. This includes a training component (our **T5.2-T5.5**) but also the idea of partnering with a journal (cf. **T5.1**). The Gene Wiki project is to run until April 2018, so we envisage organizing a joint session during at least one of the events we are both attending (e.g. during Wikimania 2016, which is to take place in Italy). Wiki4R will build on the experiences of GeneWiki.

Wikimedia Deutschland is partner in the **Commons European Training Network (CONCERTO)** proposed by the Autonomous University of Barcelona under the MARIE Skłodowska-CURIE ACTIONS Innovative Training Networks (ITN), Call H2020-MSCA-ITN-2015. This ITN, if funded, will provide a cross-disciplinary approach to research of global, local, digital and non-digital commons and initiate 15 individual research projects, including one on knowledge commons. Synergy effects between CONCERTO and Wiki4R would provide a better understanding of the mechanisms and policies supporting and governing diverse multi-stakeholder knowledge commons such as Wikidata.

The **Blue Obelisk Movement** is an informal organization of chemistry promoting and developing tools for Open Data, Open Source, and Open Standards (ODOSOS) in chemistry. Besides many cheminformatics tools, they also maintain a manually curated, CC0-waived knowledgebase of element and isotope properties which will be used in the Wiki4R pilots.

Multimedia Objects in Open-Access Publication, a recent DFG (German Science Foundation) application by the German National Library of Science and Technology (TIB) will investigate the “harvesting, indexing and provision of multimedial open access objects using the infrastructure of Wikimedia Commons and Wikidata”. If funded, this action will provide significant synergies with Wiki4R, which will be leveraged through the Leibniz research network Science 2.0, in the framework of which TIB collaborates with MfN and WMDE.

Data analysis tools like **Cytoscape** or **R** and workflow systems like **Taverna** are frequently used in research, and Wiki4R is interested in integration with these tools. Various consortium partners are involved in projects that expose data platforms in such tools. These projects include **BioVEL/ViBRANT**, **Wf4Ever**, and **eNanoMapper**, all aimed at making data integration and analysis easier.

Other relevant past and present projects are **OpenUp!** (natural history media aggregator into Europeana), **OpenAIRE** (EC Open Access Infrastructure for Research in Europe), and **Zenodo** (open access literature repository, used also by OpenAIRE).

Wiki4R will focus on use case scenarios of the VRE that cross disciplinary boundaries in that they:

1. are useful for all disciplines: scholarly publications and associated multimedia, ontology development, metadata about institutions or researchers;

2. cover a broad range of disciplines (H2020 Open Research Data Pilot, citizen science, GLAM collections, structured historical information, big data);

3. focus on information from chemistry that is used in a wide range of other disciplines (from agriculture to medicine and pharmacology to palaeontology and mineralogy to the restoration of paintings);

4. focus on information from biological taxonomy that forms the basis for research in an entire field (life sciences);

5. range from monolingual to multilingual;

6. give guidance to software development representative for a diverse set of potential users.

Ambition

The development of a completely universal, transdisciplinary, transnational, and translingual VRE is an unsolved challenge. It involves extremely complex technical, legal, procedural and social issues and is, in fact, a challenge by far exceeding the resources and funding durations typically available. The proposal is highly innovative in addressing this problem by recognizing three key points:

1. The Web itself, in the form of globally interconnected data and services, is the VRE of the future.
2. The focus must be on establishing globally accepted points of collaboration to iteratively and in an agile development process gather sufficient resources, development and training actions to provide this VRE.
3. The open knowledge community has already developed solutions and platforms for many of these issues, solutions which, with limited effort, can be brought to fruition for conventional professional research as well.

Using highly limited resources, the proposal will therefore focus on investigating and developing the functionality of Wikidata for professional scientific research. VREs built on top of existing, widely deployed and utilized platforms are rare, but this approach is essential for sustainability, trust and reaching a large community. The proposal most likely is asking for significantly less resources than many other VRE proposals in this call. In the light of the vision above, it understands itself as a pilot in a global, highly ambitious process of increasing integration, collaboration, and openness.

Professional scientists and researchers as well as citizen scientists (including "Citizen Data Scientist") will be able to use this environment. With the inclusion of Freebase into Wikidata (expected to be accomplished by the Open Knowledge community in 2015), the Wiki4R VRE will be capable of providing a unique service: for the first time, both citizen and professional scientists from any research or language community can integrate their databases into an open global structure. This point of integration can be used to publicly annotate, verify, criticise and improve the quality of available data, to define its limits, to

contribute to the evolution of domain-specific ontologies, and to make all this available to everyone, without restrictions on use and reuse. One application will be the analysis of intersections of public (governmental) and research data. An example could be the combination of disease epidemiology data (by country, by year) with public sales data of products (e.g. drugs, food) or events (e.g. concerts, movies).

Open Science in itself is an ambitious new undertaking. The proposal is ground-breaking in combining the open collaboration methods developed in citizen-driven open knowledge initiatives with the new Open Science approaches developed in professional research on a common infrastructure.

Impact

Expected impacts

The research community is experiencing a massive growth in data, a trend that is expected to continue. On the one hand, this manifests itself in the form of large data volumes and data streams with high bit rate of relatively homogeneous data. Particle physics are a prime example where these aspects of big data are dominant. These are addressed through various EU funding actions (e.g., EUDAT). On the other hand, however, the trend also manifests itself in the form of “big complexity issues”. New research information refers to previous information, often across many different disciplines. Traditional isolated databases containing verbatim information in data fields, queryable through human-oriented search portals or custom APIs, do not fulfil today’s data analysis requirements. A first improvement is to disambiguate relations between items through stable, globally unique identifiers. However, this still requires the expert knowledge in which “data silos” related data will be found. The next step in managing the complexity of interrelated scientific, societal, and governmental information is therefore to use identifiers which inform machines how to reach related information. The most general model for this is called the “Semantic Web”.

Whereas funding with respect to high data volume and high speed processing issues is ample, it may be that insufficient attention is yet given to issues of high data complexity. With the present proposal we offer a step into this direction, addressing many issues of the future of scientific data analysis in the face of high data complexity. The proposal contributes to:

- Strengthening the European Expertise in Semantic Web issues,
- interlinking research data early on,
- developing Open Science collaboration expertise in general, and especially
- developing capacities in collaboration expertise between citizen data scientists

While the Semantic Web addresses many technical issues, it becomes truly effective only if data are semantically interlinked to provide context (the final star in Berners-Lee’s five star concept). Reaching this state at the necessary scale is a complex problem. An

important answer is community data curation by an alliance professional scientists from multiple institutions as well as citizen data scientists. To achieve this, database maintainers must relinquish some control and become competent in smarter collaboration techniques.

Wikimedia projects are a prime example for crowdsourcing and massive collaboration per se, but they also have a track record of involvement in the curation of external scientific databases. Notable examples include the Gene Wiki project – which curates information about human genes on the English Wikipedia and now increasingly on Wikidata – and the Rfam/Pfam projects, which collaborate with the English Wikipedia in curating databases about RNA and protein families.

The project addresses collaboration opportunities both between professional organisations and between professional research and citizen science. An example may illustrate use cases for the first scenario:

Natural history museum collections across the world hold hundreds of millions of specimens, which are important for studying the diversity of organisms and their uses. The digitisation of theses specimens is a slow and expensive process (because of sheer numbers, diversity of objects and their preservation, hard to decipher handwritten labels, etc.). Knowing itineraries of collectors, routes and dates of expeditions can improve quality control and greatly simplifies expensive processes like geolocating specimens (essential for habitat modelling, nature conservation and climate change research). Having a transinstitutional repository of collector itineraries rather than each institution curating local data would greatly increase the efficiency of work. Existing systems like ORCID (for contemporary authors only) or VIAF do provide much information, but cannot provide enough coverage for the breadth of collectors relevant to museums. The Wiki4R VRE will provide links to ORCID and VIAF, but also support adding new persons as needed in a way, that local work can use the newly created semantic identifiers within seconds.

Highly relevant to the Wikidata argument is that the use case is in fact not limited to natural history collection. Many collectors collecting both natural and cultural history artefacts. Thus, the standard disciplinary approach to institute a shared conventional database across natural history institutions (e.g. in Europe through CETAF) is still not fully satisfying. Going even further, collections are also relevant as a historical record documenting global exploration (and often colonization) of the world and collaboration with historical researchers is highly desirable.

The Wiki4R VRE is ideally positioned for collaborative data curation across scientific disciplines, organisations, countries and languages, since it:

- is already well established and globally available to everyone,
- provides data in a structured format and in multiple languages,
- can be contributed to and corrected collaboratively, through humans and machines,
- can be linked with other ancillary data (e.g. current or historical geographical entities).

In order to realize this transdisciplinary integration potential, groundwork has to be laid by integrating domain-specific research data and scientific knowledge bases with existing Wikidata content. We will support the ontological mapping (WP2), establish tools and workflows for systematic integration (WP3) and analysis methods (WP4), and develop training materials. The effectiveness of the methods developed will be tested. Important pilot use cases are bibliographic metadata about scholarly publications and chemistry data (e.g. small molecules, biochemical pathways, enantiomers). Professional chemical researchers will closely collaborate with Wikidata's WikiProject Chemistry – a community of volunteers with a shared interest in curating chemical information on Wikidata – and professional librarians with WikiProject Source MetaData, a community interested in metadata about references cited on Wikipedia. Similar WikiProjects exist for a steadily growing variety of topics and will be our general points of contact for ensuring interoperability of our cross-domain activities with domain-specific matters. For instance, WikiProject sum of all paintings collects information about paintings in museum collections around the globe, and lists of these paintings can easily be generated on the basis of Wikidata, e.g. for Kandinsky. Other WikiProjects on Wikidata deal with music, economics, geology, medicine, or tropical cyclones.

The proposed project – in its transdisciplinary character – has also excellent potential for gender research. Existing research examples include the Wikipedia Gender Inequality Index research project, studying questions like among all the gendered biographies, which Wikipedias have the highest percentages of articles about a given gender? As the semantic data becomes richer, it will become possible to study the interrelation between gender and cultural biases, or sexuality and ethnicity biases among Wikipedias. In fact, for any recorded property in Wikidata, we will be able to see how they are biased by language.

We aim to demonstrate that the increased uptake of Wikidata as a central part a Wiki4R VRE will accelerate the pace of scientific discovery. Technically, Wiki4R will improve reliable discovery, access and re-use of public data. Not all data will be in Wikidata itself; the vision is that Wikidata becomes a hub with linking concepts and essential information, with additional detail provided as the professional research organisations. Socially, the project will lead to more effective collaboration among professional researchers and between professional and citizen researchers. Open science has great potential to increase the efficiency and creativity of research. Furthermore, true collaboration with citizens, including discussions about research goals, will greatly strengthen discussions about responsible research and innovation. The project is dedicated to building bridges between user communities. With its breadth of scope and true transdisciplinarity, the project will have impact across disciplines and can become a paradigm for structuring open data for open science.

In many ways, Open Science is an equivalent to open borders. The economic success of the removal of taxation, legislation, and other cross-border obstacles to trade in the EU demonstrates how research can gain by mastering the necessary competencies to remove obstacles between research groups, disciplines and nationally organised research organisations. However, this vision urgently needs a foundation in platforms designed for

openness, transparency and trust. Wikidata is the ideal candidate for achieving this, but critically needs enhancements to better interact with professional research.

Measures to maximise impact

Some measures to maximise impact have already previously been mentioned. Wikidata is a generic system designed for most types of structured semantic data from all scholarly knowledge domains and is available openly and widely. Of special importance is the focus on enabling re-use.

By default, any project involving Wikimedia communities or a Wikimedia movement organization will require that all project information, contents, data, and results are made available under an open license. Consequently, this project will be entirely based on open source software (Linux, Mediawiki, Wikibase, etc.) and all software developed in the consortium will be under open source licenses, maximizing the coalition that maintains the software beyond the duration of the project. Research data integrated into Wikidata likewise will be re-usable under open licenses. The project will treat data pursuant to the definition of Linked Open Data, a standard also adopted by the EU for its Open Data Portal (see also T4.1).

Reports or scientific publications resulting from the project will be published under the 'gold model' of open access, with additional archival in relevant repositories. The default license will be CC BY 4.0.

Of critical importance for the success and impact of the project is that the Wiki4R VRE becomes compatible with and integrated into the data workflows of professional scientific institutions. Interoperability will be tested with W3C unit tests (Task 4.1). Data can be created and updated both by humans and software/machine and are immediately available for further use or new semantic links (Task 4.2).

Dissemination and exploitation of results

The **exploitation plan** for the project is primarily based on the concept of openness. The value is generated as a result of lowering barriers to re-use and integration, similar to how the EU economy profits from lowering other barriers, such as tariffs, border controls, and or non-harmonized legislation. Openness is not altruism, but in many cases makes societal and economic sense.

A good example is open source software. This form of software is not the domain of hobbyist or universities alone. It makes economic sense because it allows major competitors in the IT industry to share the burden of most software development, while focusing their proprietary actions on a number of (non-open) additions.

Similarly, a global, transdomain, multilingual structured knowledge base, providing generic facts as well as research results, with open licensing and a powerful API is an excellent basis for entrepreneurial business developments. For example, Histropedia – a startup in

the UK and an Associate partner – uses Wikidata statements as the basis for location- and time-aware services that allow users to create or view timelines on topics of their choice. Media from Wikimedia Commons and related Wikipedia articles are automatically added. The system will support combining external data with globally available Wikidata-based data, to create spatio-temporal visualisations for research, education, presentations or as parts of other software applications. Intended user are academic research, education (e.g. a chart of atmospheric CO2 and average temperatures plotted against a timeline of geological events), the tourist industry, and organisations that wish to visualise their own proprietary data sets in combination with publicly available data. Another example is Music Brainz (a project of the associate partner MetaBrainz) which already closely interconnects its own work with Wikidata and which is used e.g. by the BBC (UK).

Open data can in general be an excellent basis for business. An apt example is the recent decision to dissolve the Google Freebase system in favour of the broader Wikidata integration possibilities, with Google reasoning: “they’re growing fast, have an active community, and are better-suited to lead an open collaborative knowledge base.“ We agree with Google’s assessment and consider it important to build capacities on exploiting openness such that the European economy may benefit.

Similar benefits from openness can be expected in the area of education, especially the development of Open Educational Resources (OER) and courses. The benefit will not only the availability of data that are directly relevant to the curriculum. The availability of data sets of the scale and complexity present on the Wiki4R VRE will enable courses where real research can be performed based on inquiries generated by the students rather than the teacher.

Thus, the focus of the project is on dissemination, not knowledge protection. Data present in Wikidata can be either used directly or they can be downloaded and imported into mirror deployments. Management is not through access rights, but through accountability. All the data changes are tracked and patrolled by the contributors of Wikidata. Quality control is currently mainly manual, but one of the actions of the present proposal is to develop improved automated quality control systems. Both scientists and citizens can create unit tests to provide timely quality control (Task 4.1). The results can be used to alert community administrators as well as professional research organisations curating their data on Wikidata.

Sustainability

The **sustainability** of the Wiki4R VRE critically depends on the financial sustainability of Wikidata. Wikidata is a project of both the global Wikimedia Foundation and Wikimedia Deutschland (Germany). Since 2012, both organisations have invested significant amounts of their own funds, supported by third-party funding from, e.g., the Allen Institute for Artificial Intelligence, Google, the Gordon and Betty Moore Foundation, and Yandex. Wikidata has already become a core technology for the functioning of Wikipedia and is thus essential to the Wikimedia movement. It is supported by a number of additional stakeholders from the cultural, information technology and educational arenas who see its

potential for making content and knowledge accessible. Wikidata's business plan for financial sustainability includes a diversity of financial and human resources which are currently secured or under development:

- The invaluable human resources provided by the community of Wikidata online volunteer contributors and editors (currently over 14,000 individuals and growing fast)
- Base funding from the budget of the Wikimedia Foundation (which raised USD 37 million in online donations in the 2013-14 fiscal year)
- Base funding from the general operating budget of Wikimedia Deutschland (which raised USD 8 million in online donations)
- Funding through partnerships with corporate donors
- Funding from private foundations (WMDE fund development staff is currently working with a strong focus on fund acquisition from European-based foundations with an interest in free and open knowledge)
- Public educational and cultural institutions and entities of government, interested in partnering with WMDE and its broad community of online volunteers.

A direct monetization of Wikidata is neither intended nor possible, as it conflicts with core principles of the Wikimedia and Open Knowledge movements. It is, however, possible that at some point computationally expensive services may be provided at a charge or under dedicated funding by public or private partners depending on these services. In the framework of the current proposal, we consider a testing of such services to be premature.

Communication activities

As a basis for engaging with the target communities and beyond, all communications will be as open as possible, as is common practice in both Open Science and Wikimedia contexts.

That practice was already applied to the drafting process of this proposal itself, which was completely open from the initial blog post to the outline on Wikidata and from the actual drafting in public Google docs to the archiving of the final version under a Creative Commons Attribution license in a dedicated Zenodo community and a dedicated page on Wikidata for post-submission updates.

This was complemented by daily communication via a public mailing list, and a dedicated Twitter hashtag as well as contributions to the Wikidata mailing list and community newsletters related to Wikidata or cultural partnerships, in addition to public hangouts and the various communication channels on Wikidata itself.

One article about the project was viewed well over 6000 times prior to proposal submission , and another one offered "uncritical cheering" for the project. Judging from such feedback and from the contributions we received from the community during the drafting process, this multi-channel interactive approach is fruitful. We were further encouraged by public

attention being directed to the issue of open grant proposals in response to one of our publications.

A final underpinning of our communication strategy is the awareness that Wikidata was conceived as a tool to help manage data for Wikipedia. While Wikidata has since grown more of an identity of its own, this aspect is still important, and it provides exposure at scale and in multiple languages.

On that basis, the detailed communication strategy varies depending on the characteristics of the respective communities:

Individual professional researchers will be reached through the networks of the project partners, through publications, mailing lists, training events and through conference attendance. We will also address the lack of professional recognition for community curation by exploring ways to couple it to formal publications. For example, the Gene Wiki Reviews couple the publication of review articles on specific genes with contributions to the corresponding articles on the English Wikipedia. This follows in the footsteps of the journals RNA Biology and PLOS Computational Biology, which have been coupling Wikipedia pages and review articles in a similar fashion. One of the MfN team members (DM) is involved in the PLOS initiative, and associate partner Pensoft (who is the publisher for several MfN journals) is interested in such an approach. While such initiatives have created a relatively low number of articles so far, they are widely known in the respective communities and can thus serve to raise community awareness of our activities.

Professional Research organisations: All project partners are involved in professional networks, which will be leveraged to reach out to professional organizations outside the project. For instance, MfN and WMDE are both members of the Leibniz research network Science 2.0, and UPM acts as the Spanish node of the DBpedia network.

Furthermore, we will leverage the growing network of interactions between the Wikimedia and research communities, which has so far been focused on Wikipedia, Wikimedia Commons and to some extent Wikisource, but is now extending towards Wikidata.

One particularly successful initiative in this framework is that of a Wikimedian in Residenc e, i.e. an active member of the Wikimedia community working inside an organization or institution on enhancing the interaction between the two. Such Wikimedians in Residence have been active in about 100 institutions around the globe over the last 5 years, and they form the nucleus of a very active community at the intersection between Wikimedia and the cultural sector, which increasingly extends into research-related organizations like the Swedish Agricultural University, Cancer Research UK, or ORCID. The Royal Society of Chemistry, specifically, is an Associate partner and will assist the project in engaging the chemical community.

These cultural partnerships have also led to technical developments, e.g. several Wikimedia chapters partnered with Europeana to create a toolset that facilitates the upload of images and other media from heritage collections to Wikimedia Commons with proper metadata.

Higher Education: The consortium has four universities as members (UPM, UM, UOC, UPS), three of which will collaborate to organize a MOOC (T5.4), and all partners are involved in the development of course materials (T5.3). The non-university partners also have strong ties to higher education, e.g. doctoral research being performed at MfN, Europeana materials being used in lectures, and OER conferences organized by WMDE. Furthermore, many users (both contributors and readers) of Wikidata are students or Early-Stage Researchers, and hundreds of courses are organized around the globe each year that have a Wikipedia component and that increasingly involves Wikidata as well.

Citizen Science projects: The European Citizen Science Association, the secretariat of which is hosted at the Museum für Naturkunde Berlin, has established important communication channels to a large number of citizen science projects in Europe. We will communicate through these channels (web site, newsletters, events and meetings).

The **Wikidata community** primarily communicates through the well-established and widely used Wikidata community portal. It facilitates introduction of new users and editors to the tools, rules and practices, enables users to submit requests and engage in topical on-wiki discussion. It further provides the platform for Wikiprojects, which are groups of editors working together to improve Wikidata. Wiki4R will utilize the existing channels of this platform to provide updates, resolve integration issues, and introduce new users from the science communities to the world of Wikidata.

Wikimedia communities in general: Wikidata is the fastest-growing Wikimedia project globally in terms of the increase in number of volunteer contributors. However, there are other Wikimedia communities – including Wikipedians, Wikimedia organizations, and contributors to cultural heritage projects – whose members have a vested interest in the development of Wikidata. Communication activities for these people and groups will use well-established channels as well, including a variety of Wikimedia blogs, online newspapers, Meta-Wiki pages and Twitter.

Policy and decision makers: Outreach and continuous communication to this group will utilize, among other channels, the joint Brussels office of the EU-based Wikimedia Chapters. Also known as the Free Knowledge Advocacy Group EU (FKAGEU), this Alliance of Open Knowledge advocates will be able to directly communicate policy-relevant project results, barriers, insights and relevant project outputs through its strong network of contacts with Members of the European Parliament and other key decision makers. Some of the main policy priorities of the FKAGEU are EU copyright reform, free and open access to public works (which includes scientific articles produced by publicly funded researchers) and freedom of panorama. All of these policy issues are of crucial importance to the goal of increasing the number of items and data sets that are available under open license and linked in Wikidata.

Implementation

WORK PLAN

Overall structure of the work plan

The five work packages (WPs) can be briefly characterized as follows:

WP 1: "Coordination and management" will provide the administrative support for effective project management and support the organisational structures and governance mechanisms for the efficient coordination of the Wiki4R consortium. It will facilitate communication and exchange of information between consortium partners, WPs, and organisational bodies, including an independent Advisory Board. Main outputs will be the implementation and updating all documents required for the full implementation of the projects, including the management of deliverables and all interim and final financial and scientific reports. Another main output will be a highly transparent and open communication, which provides new avenues for assessing and engaging with the research process and its outcomes.

WP 2: "Semantic modelling" will work on community-agreed profiles for the most relevant semantic concepts (core classes and core properties) in a selected number of scientific domains. Profiles include semantic mappings as well as associated policies and best practices that meet the functional requirements of the research use cases addressed in WP3 and WP4. Furthermore, this WP will match these classes and properties to ontologies used by the scholarly community to allow integration of Wikidata into research data, using the common ontology used by both. Main outputs will be property profiles required for the classes of data sets to be integrated in Wiki4R as well as for common classes needed for the research context. Policies and best practices will be documented and semantic mappings provided where each class and property will have at least one relation to external ontologies. Multiple mappings may be provided where required (especially if different use cases have different needs).

WP 3: "Integrating research resources with Wikidata" will demonstrate how research data can be integrated with Wikidata in the Wiki4R VRE. For this purpose, it will incorporate a selection of external research resources into the VRE, demonstrating pilot workflows for semantic integration, quality assurance mechanism, and bidirectional information flow. Attention will be given to the mechanisms in which Wiki4R can promote the open licencing of external resources that are yet unavailable for re-use. Main outputs will be workflows, best practices documentation, approaches to measure the quality of data, and foremost working examples that serve as paradigms and motivations for future applications. In addition, resources suitable for the future application of Wiki4R will be identified and the motivation and pitfalls around sharing data experienced in the pilots will be summarized in reports. This WP depends on WP2.

WP 4: "Enabling the use of Wikidata in research contexts" will integrate Wikidata into the daily workflows of professional scientists to create a full VRE. To address confidentiality issues of ongoing research projects, this WP will develop solutions to dynamically mirror the content of Wikidata and provide controlled information interchange using Semantic Web technologies. Scientists will be able to curate information in Wikidata using workflow-oriented dedicated editing interfaces that will interact with Wikidata via its APIs. The approaches will be tested in the laboratories of one of the project partners, using citizen science projects, and in a collaboration with Europeana. This WP depends on WP3.Main outputs will be a new interaction between Wikidata and a SPARQL endpoint for complex queries, and a ready-to-use virtual machine that allows hosting a Wikibase software instance compatible with Wikidata and Wiki4R in a protected institutional network. Further output will be solutions developed around the use cases to test the approach and reports on the effectiveness of Wikidata use in these use cases.

WP 5: "Dissemination, stakeholder engagement and training" will disseminate the outputs generated by the project to communities of stakeholders. It will practice openness through open communication to the extent that this is practical. In combination with the multiple, largely bidirectional communication channels outlined in the communication strategy, this allows to engage diverse communities more profoundly than traditional methods. Main outputs will be communication and engagement activities, including training activities directed at science students, professional researchers as well as members of the Wikidata community. The ultimate output will be capacities in the European Research Area in engaging in Open Science, using Wiki4R as a hub for research.

Timing of the different work packages and their components

The detailed timeline for implementing the Wiki4R work plan, for all Work Packages and individual tasks including the timing of deliverables can be obtained from Fig. 3 (Gantt chart).

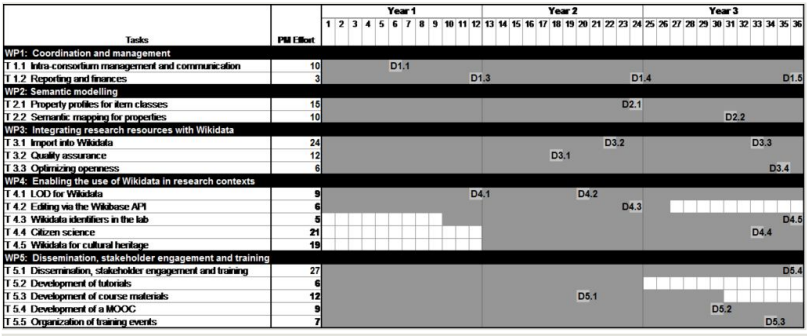

Figure 3.

Timeline of work packages and their deliverables.

Detailed work description

Work package 1 - Coordination and management (Table 2)

Table 2.
WP1 - Coordination and management

Work package number	1	**Start Date or Starting Event**					M1
Work package title	**Coordination and management**						
Participant number	1	2	3	4	5		
Short name of participant	**MfN**	UPM	UM	WMDE	UOC		
Person/months per participant:	9	1	1	1	1		

Objectives

The objectives of this work package are to oversee administration, operational management, and overall implementation of the project, including internal effective communication and collaboration between the coordinator, individual consortium members, and the European Commission (**T1.1**), to organize and administer consortium networking and governance, including supporting relevant bodies and meetings, including Advisory Board (AB), ensuring efficient and effective management and decision-making procedures (T1.1) and to implement sound financial management as well as controlling systems and quality assurance of deliverables and periodic and final Reports (**T1.2**).

Description of work

Task 1.1 Intra-consortium management and communication (**Lead MfN**; all partners; Month 1-36)

The core part of this task will be the organization and coordination of the consortium bodies and their meetings, particularly meetings of the Steering Committee (monthly remotely), and the General Assembly (once a year in person), as well to ensure for efficient and regular communication between the coordination, the consortium bodies, and all partners. In person meetings will be aligned with independent community meetings (such as the “Wikimania”, the yearly meeting of the Wikipedia/Wikimedia/Wikidata community). This task will further establish a high level scientific Advisory Board. The Advisory Board will help to monitor progress, and provide independent guidance and advice to the coordinator, Steering Committee and the entire consortium. Advisory Board members will be supported in their function by the project logistically and with limited secretarial functions. The coordination office will also provide for direct links between the Advisory Board and all consortium bodies, groups and partners as needed. Daily communication will primarily happen via electronic means, including email, Skype and the #wikidata channel on the irc.freenode.net network.

Task 1.2 Reporting and finances (**Lead MfN**; all partners; Month 1-36)

This task includes the preparation and coordination of the periodic scientific and financial project reports as requested by the EC, and of the Data Management Plan according to the H2020 Open Data Pilot. Further activities include the coordination of common administrative tasks, such as elaborating and providing templates for detailed planning and reporting of the tasks in each WP; providing formats for compiling and editing progress reports; monitoring progress through reaching milestones and deliverables; quality control and submission of products and deliverables. Financial monitoring will ensure a transparent financial distribution of the EC grant, including setting and monitoring payment schedules, and terms of reimbursement. We will also prepare financial controlling reports for the overall monitoring of the project.

Deliverables

D1.1 Data management plan (Month 06)

D1.2 Periodic report M12 (Month 12)

D1.3 Periodic report M24 (Month 24)

D1.4 Periodic and final report M36 (Month 36)

Work package 2 - Semantic modelling (Table 3)

Table 3.
WP2 - Semantic modelling

Work package number	2	**Start Date or Starting Event**				M1
Work package title	**Semantic modelling**					
Participant number	1	2	3	4	5	6	7
Short name of participant	MfN	**UPM**	UM	WMDE	UOC	EF	UPS
Person/months per participant:	3	6	6	6	1	2	1

Objectives

The goal of this WP is to assist the Wikidata community in establishing and setting up a framework of classes and properties (by using the W3C Web Ontology Language, OWL) to enable the support of the selected interdisciplinary VRE research fields (**T2.1**). Specifically, it develops profiles of classes and properties for concepts the project first focuses on, such, isotopes and their properties, like decay, research papers and properties, like its DOI, title, authors list (using the ORCID), and metabolites and other small molecules and properties like their chemical structure, boiling and melting point, and external database identifiers. The second task (**T2.2**) will focuses on the required mappings between Wikidata elements (i.e. classes and properties) and elements in external ontologies. Where needed, recommendations will be made to refine or add new

classes and properties to allow easier or more accurate integration with research data. Besides actual mappings, also guidelines and procedures will be established that streamline the process of building and extending class and property vocabulary and of populating items with statements involving those properties. Here, the multilingual nature of Wikidata will be included. We will reuse community adopted ontologies and vocabularies, such as ChEBI and CHEMINF, and standards like the CIDOC Conceptual Reference Model used in cultural heritage documentation.

Description of work

Task 2.1: Recommendations for classes and their property profiles (**Lead: UM**; all partners; Month 1-36)

This tasks focuses on the development of VRE-specific property profiles for classes of Wikidata items for a number of disciplines. Profiles are defined as sets of properties and associated policies and best practices that meet the functional requirements of specific research use cases (**WP3** and **WP4**). Profiles are not limiting the application of properties which are absent from the profile; rather they guide the interaction by reducing the mapping process to a manageable set of properties. Where possible, existing community efforts towards creating property profiles will be built on. For instance, generic lists already exist of properties available for creative works or periodicals, as well as draft recommendations for scholarly articles and chemicals. However, properties considered for profiles are not limited to properties that already exist on Wikidata; new ones will have to be introduced. An important part of this task is also to work with Wikidata development (partner WMDE) to increase the range of available data types (e.g. for supporting units or geo shapes). Property profiles will be created for a significant number of classes that fall under the use cases outlined in the “Concept and approach” part of the Excellence section of the proposal. The focus will be on the 50-100 most widely used DBpedia classes as filtered by our use cases. Priority will be given to profiles required for data sets to be integrated in T3.1, including form project partners (e.g. minerals, meteorites and taxa in the case of MfN’s collections).

Task 2.2: Semantic mapping for properties (**Lead: UPM**; all partners; Month 1-36) This task addresses the need to map the select classes and properties from **T2.1** employed in Wikidata to the larger semantic web world, including DBpedia. We will identify ontologies and aim at releasing the needed terms as a single "Wikidata for research" ontology. Example options include the ChEBI and CHEMINF ontologies for the chemistry domain (for the life sciences the BioPortal and Ontology Lookup Service are useful). Properties used in property profiles of target item classes from **T2.1** will thus be mapped across multiple established ontologies relevant to the respective VRE use case. On that basis, a pattern will be built that allows statements about the respective item classes to be expressed in RDF. Mappings will be provided in a machine readable way, to aid the RDF export of Wikidata as outlined in **T4.1**.

Deliverables

D2.1 Report on property profiles (Month 24)

D2.2 Report on semantic ontology mappings (Month 31)

Work package 3 - Integrating research resources with Wikidata (Table 4)

Table 4.
WP3 - Integrating research resources with Wikidata

Work package number	3		**Start Date or Starting Event**				M1
Work package title	**Integrating research resources with Wikidata**						
Participant number	1	2	3	4	5	6	7
Short name of participant	MfN	UPM	**UM**	WMDE	UOC	EF	UPS
Person/months per participant:	12	5	13	7	1	3	1

Objectives

This work package is concerned with external research resources with the VRE for the use cases in the focus of the project. This includes identifying external research resources suitable for integration with Wikidata, establishing and demonstrating workflows for such integration (**T3.1**), establishing quality assurance mechanisms (**T3.2**), facilitating the reuse of Wikidata, and analysing the conditions under which external resources are being made openly available in a way that is suitable for integration with Wikidata (**T3.3**). The tasks focus on both defining the requirements as well as the implementation of these tasks, and a selection of Open Data resources will be integrated to demonstrate how the process of this integration practically works, with the benefit that Wikidata grows in size and becomes even more useful to the research community. It builds on existing and well-documented experiences with such data integration, e.g. in the context of authority control or human genes, or from Wikipedias.

Description of work

Task 3.1: Integrate external databases with Wikidata (**Lead UM**; Month 1-36)

This task is concerned with enriching Wikidata with CC0 data available from external sources, taking into account the property profiles and the semantic mappings created in **T2.1** and **T2.2**. Work here will begin with the integration of one cross-domain database (The European Library) and domain-specific ones (Blue Obelisk Data Repository). Integration will be implemented as Open workflows, for which a suitable platform will be selected. A good candidate here would be the Open Science Framework (OSF) hat is being developed by the Center for Open Science (COS), an Associate partner. The task consists of the following subtasks:

European Library Bibliography: The European Library aggregates bibliographic data across European libraries and makes the data available under CC0 through RDF and an API. While current data may focus on bibliographic information about articles, books, etc., citation of other types will be considered too, like software and data citations, for which we aim at full compatibility with the DataCite Metadata Schema. (Subtask Lead EF)

The **Blue Obelisk Data Repository** (BODR) is a CC0 data collection of chemical element and isotope information collected and maintained by the Blue Obelisk movement. The data is manually collected and curated from primary sources from, among others, the IUPAC and scholarly publications, which could be listed accordingly. Classes involved here include chemical elements, materials consisting of a single element (e.g. diamond and oxygen gas), and isotopes. Properties for these classes include element symbol, melting points, decay constants, etc. The goal of this task is to make a BODR release based on Wikidata content. Collaboration with the “Groupe de Chimie Analytique de Paris Sud” is anticipated as well. (Subtask Lead UM)

(Human) Metabolites: For chemical structures, human metabolites will be a focus area. The Recon2 project created systematic model of the human metabolism under the CC0 waiver. This will feed back into the semantic modelling of chemical compounds, as per **T2.1** and **T2.2**. Central here are chemical compounds, metabolites, with property names, and identifiers. Challenges include the charge states of metabolites in this data set and linking those to the uncharged species. This work is complementary to but partially overlaps with the work by the Gene Wiki project, which focuses on drug compounds. Collaboration with WikiPathways is anticipated as well. (Subtask Lead UM)

Other data sources: The task will then expand towards the integration of other data sources, possibly including MusicBrainz, research-related subsets of Freebase, a subset of PubChem, the H2020 Open Data Pilot and further resources that are to be identified in **T3.3**, including from MfN and Europeana. Options here are the integration of information about institutions and to support the H2020 Open Data Pilot. In this pilot, considerable amounts of data will be made publicly available over the course of the project, especially in those key areas where the pilot is mandatory, such as the NanoSafety Cluster community. A subset thereof (as identified through **T3.3**) will be suitable for integration with Wikidata, which will be handled in this task.

Task 3.2 Quality assurance (**Lead: UPM**; all partners; Month 1-36)

This task is concerned with systematically monitoring and improving the quality of the information already available on or newly added to Wikidata. This includes the handling of provenance information (in conjunction with **T3.1**), along with provisions for data citation (in conjunction with the Data Citation Principles and the property profiles developed in **T2.1**). For instance, one of the reasons for the growing reference rot in scholarly publications is that the URLs of databases change over time. Identifier systems like Digital Object Identifiers (DOI) or the Handle System address this issue, but not all databases use them now, fewer have used them in the past, and not all references to those databases having such identifiers make use of them. In Wikidata, this kind of information could be modelled

by making statements using properties like official website or URL formatter with qualifiers that indicate start and end times. Tools such as those developed in **T4.3** can then reason upon that information and turn a dysfunctional data citation back into one that points to the current location of the data if the database still exists.

The task also includes establishing mechanisms for community editors to review and verify existing and incoming information, both on a factual and linguistic level, as well as tools to assist with that. Furthermore, it includes measures to assess and increase the consistency of information on Wikidata. This involves diagnosing erroneous statements (e.g. related to the death of a person), which may occur because external sources conflict, or information across languages, or because related statements are incomplete. For statements identified this way, we will develop measures to fix them on a programmatic basis or through interaction with Wikidata editors, while leaving room (as the Wikidata data model allows) for inconsistencies that simply reflect inconsistencies outside Wikidata, such as territories being claimed by more than one independent country. Finally, the task includes ways to propagate those fixes (or other annotations of the detected inconsistencies) to external databases, to alert original sources if appropriate, and to notify interested Wikipedia editors or WikiProjects that such inconsistencies were detected and acted upon.

Task 3.3 Optimizing openness (**Lead: MfN**; all partners; Month 1-36)This task consists of two main activities: (1) the identification of sources of open data suitable for integration into Wikidata (public domain or CC0 waiver), (2) an analysis of the motivations for sharing such data openly, with the aim of informing future open-data policies at project partners or more generally, including for H2020.

(1) Only a small subset of the publicly available research data is suitable for integration with Wikidata, with legal, ethical (privacy, etc.), and technical interoperability being major factors, along with considerations of quality, maintainability and scope. For instance, while many datasets that would otherwise fit into Wikidata are not free of reuse restrictions, some datasets like the metadata in Europeana are available under CC0 but do not fully match the scope of Wikidata. Additionally, while efforts to make data repositories more visible have been stepped up in recent years, data discoverability remains a challenge. We will generate an overview of potential data sources for Wikidata and identify datasets suitable for import or mapping within the framework of the project, as per **T2.1** and **T3.1**. This overview will include general repositories like re3data as well as data sources in the respective fields (both domain-general like BioSharing and more specific ones) and from citizen science projects (in conjunction with T4.4) and the H2020 Open Data Pilot, taking into account that its requirements on openness are not as strict as those at Wikidata.

(2) Building on that survey of data sources, we will analyse underlying motivations for sharing data openly, in order to distil best practices and provide recommendations, conscious of existing barriers to sharing. The sharing of data and other resources is an integral part of research endeavours. In the Web age, most new research objects are digital, and many legacy ones are being digitized. Once digital, they can be easily shared over the Web, and from there, it is technically only a very small step towards opening them up for reuse by a potentially global and cross-disciplinary audience. Socially, this step is

significant, and few incentives beyond altruism exist for institutions, research groups or individuals to fully embrace openness. Mid- to long-term effects of openness have been the subject of prior investigations that established, for instance, citation advantages for open access articles or for publications associated with open data or open source software. The situation is much less clear for immediate and short-term benefits, but if such benefits exist, an analysis of best practices around them can help harness their potential for data providers, the wider research community, and society at large. As one of the founding signatories of the Bouchout Declaration for Open Biodiversity Knowledge Management (along with associate partners Plazi and Pensoft), MfN has a special interest in these issues, and as the coordinator of this project that is centred around data sharing, a special responsibility.

This activity is thus concerned with identifying benefits that accrue to those who share their research openly, as laid out in the Bouchout Declaration, but not limited to the field of biodiversity research. For instance, it has been suggested “that data sharing, especially sharing data through an archive, leads to many more times the publications than not sharing data.” Similarly, institutional data sharing has been associated with “becom[ing] a canonical reference point.” We will also analyse potential pitfalls associated with not sharing data openly, again a case that could help build momentum behind the Bouchout Declaration and similar initiatives. For instance, many paintings from heritage collections are not available online through the respective institution but in a multitude of variations from other sources, to the point that it becomes next to impossible to discern properties that these digital representations may have in commons with the original. In those cases, the above-mentioned reference point is missing. The wider societal framework in which institutions operate may affect openness too, e.g. through legislation, trade negotiations, or a competitor’s openness. We will take this into account in providing recommendations for data sharing policies, focusing on the institutional and European levels. A final role of the survey is to highlight barriers to the use of Wikidata in research contexts, to suggest measures to lower or overcome those barriers within the framework of the use cases, and to identify data sharing communities whose curation workflows might benefit from integration into Wikidata, as prototyped by Encyclopedia of Life or WikiPathways and addressed in **T4.4**.

Deliverables

D3.1 Report describing quality assurance measures (M18)

D3.2 Lessons learned from integration of TEL, BODR, and Recon2 (M22)

D3.3 Report on other integration uses cases (M33)

D3.4 Report on the motivations and pitfalls around releasing open data (M35)

Work package 4 - Enabling the use of Wikidata in research contexts (Table 5)

Table 5.
WP4 - Enabling the use of Wikidata in research contexts

Work package number	4		**Start Date or Starting Event**			M1	
Work package title	**Enabling the use of Wikidata in research contexts**						
Participant number	1	2	3	4	5	6	7
Short name of participant	MfN	UPM	UM	**WMDE**	UOC	EF	UPS
Person/months per participant:	13	4	4	14	2	10	13

Objectives

The aim of this WP is to connect research workflows directly with the Wikidata knowledge base (**T4.1**). To this end, research laboratories will collaborate directly with the Wikidata community of contributors (**T4.2** and **T4.3**). A platform will be created, where research laboratories or organisations can work on data partly in a private environment and partly in the shared open environment, while at the same time increasing the amount of work that occurs as open science. The interface for direct work of scientists on Wikidata will be improved with the goal to curate information directly in lab tools. A better synergy between citizen scientists (**T4.4**), professional scientific researchers and professional research organisations will increase the quality of the data available for all (**T4.5**). As examples, the tasks will help to plan the future use Wikidata as a VRE in research organisations.

Description of work (**Lead WDME**; Month 1-36)

Task 4.1 Wikidata as Linked Open Data (**Lead UPM**; UPS; Month 1-36)

This task will provide a data interchange mechanism between Wikidata and Wiki4R tools. The classes and properties will initially be those defined by Wikidata and later those added in **T2.2**. This mechanism will be used to provide a query endpoint (based on SPARQL), such that a data curator can change information on Wikidata and query and reuse it via SPARQL in Wiki4R. During this task, we will develop:

- Export information from parts or all of Wikidata as RDF in near real time (as fast as technically achievable) (UPM)
- Provide a SPARQL endpoint (UPM)
- Establish a mechanism for syncing the information between multiple Wiki4R instances (see Fig. 1, first specification UPS + UPM / implementation & deployment UPM)
- Package this Wiki4R mirroring mechanism in virtual machines (at least one instance in the Wikimedia Labs' cloud and one in the UPS cloud) and check their interoperability using Tests For TripleStores
- Support researchers in creating their own unit tests, to check the quality of the data automatically (see **T3.3**) and establish a continuous delivery workflow (UPS)

The aim is to create a standard workflow for the quick installation of a Wiki4R SPARQL endpoint in an institutional cloud where:

- the data can be reused without performance or interoperability problems,
- errors and inconsistencies in Wikidata's data can be detected (cf. T3.3), and
- information can be updated on Wikidata at least manually (errors, additions)

Task 4.2 Editing Wikidata via the Wikibase API (**Lead UPS**; all partners; Month 1-26)

Building on the mirroring mechanism developed in **T4.1**, the scientists' own tools will allow them to edit via the Wikidata API and reuse the updated information directly from within their research workflows. The task will build on an existing OAuth scheme for remote editing of Wikidata . It will be configured such that relevant Wikidata policies are honoured, without requiring researchers to know every detail of these policies. Special attention will be paid to proper provenance and data citation (in collaboration with **T3.2**). The development of research tools where scientists can edit Wikidata from within their workflows will be tested by scientific researchers using a large number of chemical analytical methods and techniques.

Task 4.3 Wikidata identifiers in the lab (**Lead UPS**; all partners; Month 10-28)

This task will integrate the functionality of an existing laboratory database portal developed by UPS into Wiki4R to provide a test case for Wiki4R. It will inform the design of Wiki4R research workflows and of applying Open Science and Linked Open Data principles on the basis of Wiki4R. Use cases, from across the laboratories of UPS, in which research workflow tools require semantic identifiers for materials, procedures, instruments or other facilities, will be explored. A major goal is to enhance the reproducibility and transparency of research by integrating a semantic documentation. One use case in analytical chemistry will be selected to be prototyped. In conjunction with outreach activities and building on the results of T3.3, the final result will be a report on the place of Wikidata in science and the role of science in Wikidata.

Task 4.4 Citizen science (**Lead MfN**; MfN; Month 13-36)

The Wiki4R VRE is uniquely positioned to serve citizen science projects. This task will explore ways to (1) involve the Wikidata community with the curation of scientific knowledge bases, (2) connect Wikidata to external crowd collaboration projects with a scientific focus.(1) We will extend the community curation approaches employed by the Gene Wiki and Rfam/Pfam projects to other areas: building on the outcomes of **WP3** (**T3.3** in particular), as well as on past collaborations (e.g. with the Encyclopedia of Life and WikiPathways), we will work with a small group of professional research communities that are already sharing their data openly to explore how their curation workflows might benefit from a citizen science component integrated with the Wikidata platform and community. This will include MfN groups along with the laboratories partnering in **T4.3** to explore the use of Wikidata identifiers in the lab, and communities using Wikibase in their curation workflows, e.g. the EAGLE project. (2) Starting with citizen science projects identified in **T3.3** as suitable for integration with Wikidata, we will work on integrating the workflows on

those platforms with those on Wikidata. The work is based on appropriate property profiles (**T2.1**), mappings to suitable external vocabularies or ontologies (**T3.1**), and quality assurance mechanisms as identified in **T3.2**.

To illustrate how this might look like, consider the classical case of a curator on Galaxy Zoo being presented an image of a galaxy and tasked with the assessment of whether that galaxy is elliptical, spiral or neither. Depending on the assessment of that curator or of multiple curators on Galaxy Zoo, the corresponding Wikidata item for that galaxy could then be created or updated as to whether it is an instance of an elliptical or spiral galaxy. At the same time, other metadata about the galaxy could be imported from Galaxy Zoo or its sources, while the image that the Galaxy Zoo curator had seen would go onto Wikimedia Commons, with automated classification on the basis of that metadata, annotations of the objects visible in the image, and a statement on that galaxy's Wikidata item that the galaxy is depicted in that image on Wikimedia Commons. Subsequent rounds of galaxy classification – e.g. targeted at students of astronomy or image analysis – could then be built to decide which subtype of elliptical galaxies that particular galaxy belongs to (e.g. cD galaxy).

Such classification tools have already been prototyped as Wikidata games, allowing for instance to annotate Wikidata items about species with corresponding images from Wikimedia Commons, or items about people with statements about their gender, or items about books with statements about their authors. These prototypes can in principle be played on both desktop and mobile devices, and they produce statistics that could be fed into the calculation of altmetrics. We will work to develop these prototypes further for a subset of our use cases (e.g. animal migration, or identification keys for minerals or animal sounds), closely align them with the quality assurance measures proposed in **T2.3** and integrate their use into the outreach and training activities in **WP5**, especially the tutorials developed in **T5.3**, the MOOC in **T5.5**, and the training events in **T5.7**.

Task 4.5: Wikidata for cultural heritage (**Lead EF**; MfN; Month 13-36)

This task will explore ways in which researchers in cultural heritage (and digital humanities in general) and the Wikidata community can better interact in the aggregation and curation of cultural heritage information. Using focussed subsets of data from Wikidata (through WP3), Wikimedia Commons, and Wikisource, the task will engage with researchers within the digital humanities to define use cases. The present coverage and quality of Wikidata content for these use cases will assessed, and, building on existing Wikidata tools and metadata games, new metadata games supporting rich semantic annotation will created. Towards the end of the project, the use of the Wikidata in different research scenarios will be easier, so as to provide a good basis for wider adoption across the cultural sector.

The subsets will focus on audio recordings, historic newspapers, and World War I, themes, where the collections provided by the Europeana Foundation have excellent depth. Audio recordings are an important topic at the MfN, which houses one of the largest animal sound archives, and are at the core of activities of MusicBrainz, an associate partner of the project.

Deliverables

D4.1 SPARQL endpoint for Wiki4R (Month 12, lead UPM)

D4.2 Mechanism for semantic knowledge-base mirroring (Month 20, lead UPM)

D4.3 Research-oriented Wiki4R edition of Wikidata software (Month 24, lead UPS)

D4.4 Report on usability of Wiki4R in citizen science and cultural heritage research (M. 33 MfN)

D4.5 Report on Wiki4R integration and usability tests in science laboratories (Month 36 UPS)

Work package 5 - Dissemination, stakeholder engagement and training (Table 6)

Table 6.
WP5 - Dissemination, stakeholder engagement and training

Work package number	5	**Start Date or Starting Event**				M1
Work package title	**Dissemination, stakeholder engagement and training**					
Participant number	1	2	3	4	5	6	7
Short name of participant	MfN	UPM	UM	WMDE	**UOC**	EF	UPS
Person/months per participant:	11	4	12	8	15	5	5

Objectives

The objectives of this work package are to disseminate not just the results but the process behind building and using this VRE to communities of relevant stakeholders, especially the research and Wikidata communities (**T5.1**) and to engage with these stakeholders around using Wikidata as a Virtual Research Environment and how to collaborate within that framework (open licenses, WikiProjects, cross-cultural issues), and to provide training for that (**T5.2**-**T5.5**).

Description of work

Task 5.1: Dissemination and community engagement (**Lead MfN**; all partn.; Month 1-36)

Each partner will, within this task and over the project lifetime, actively contribute to the publication of articles and to presentations in seminars, workshops and conferences, based on project rationale, methodology, and outputs of the project. We will also use social media to that effect. Besides these classical dissemination channels, the open nature of the project allows to offer an additional and more profound layer of dissemination, one that conveys not just the results of the project, but the process behind achieving them, and even the option of contributing. This applies to content, software, data, policies, reports and publications developed in the framework of this project, and their respective identifiers.

Nurturing communities of users is one of the key challenges for VREs. We will address this by a) engaging with the existing Wikidata community about research-related matters, b) engaging with the research community about Wikidata-related matters and c) bringing the two together. For example, UPS will organize special days where their scientists will learn how to use wikis – and Wikidata in particular – for collaboration. These events will also be used to test the user-friendliness of the platform, in conjunction with **T4.3**.Inspired by existing attempts to couple journal articles and Wikipedia contributions (see description of the Gene Wiki), we will also explore ways in which something similar could be achieved with Wikidata. The three journals published by MfN would be a candidate here, as would the Biodiversity Data Journal published by associate partner Pensoft.

Task 5.2: Development of tutorials (**Lead UOC**; all partners; Month 1-24)

Development of different tutorials for WP2 (semantic mapping), WP3 ("Integrating research resources with Wikidata") and WP4 ("Enabling the use of Wikidata in research contexts") as well as on the benefits of and barriers to openness that will be identified in T3.3. These tutorials should be available online under terms compatible with reuse on Wikimedia platforms. They would focus on text and video recordings but also make use of tools like the Wikidata games to inform about key features of the site and its operation. They will both provide new training resources for participants and improve already existing documentation on Wikidata.

Task 5.3: Development of course materials (**Lead UOC**; all partners; Month 1-30) The goal of this task is to develop freely available materials that can be used in university courses and on Wikimedia platforms, so future skilled workers will make best use of the open-data infrastructure provided by Wikidata and other organizations. The intention is to provide materials for a number of different data usage scenarios. This might range from simple data retrieval to more complex treatments of data-sets like GIS and machine learning. The material will be available online, e.g. via Wikiversity, where UPS has organized such courses in the past. It will be kept up to date, and translated into at least two European languages in addition to English. It will be suitable for injection into graduate and postgraduate coursework, but also suitable for on-the-job-training. This task will be carried out in conjunction with WP2 (mapping), WP3 (integrating resources via Wikidata) and WP4 (Wikibase tools), since materials will be mostly designed from their respective outcomes. Teaching materials will be mostly in text form, adding some educational treatment and providing practical exercises and self-assessment tools. Specific tasks will be: compilation of source materials, writing, adding educational treatment, reviewing materials by consortium partners as well as the Wikidata community, editing, translation, and open publishing.

Task 5.4: Development of a MOOC (**Lead UOC**; all partners; Month 1-36) Development of a MOOC (Massive Online Open Course) that will be based on Wikidata, tightly integrated with other Wikimedia platforms, and include different learning paths according to the background and interests of participants. Specific steps in this task are: a) gathering and adapting existing resources about Wikidata and about our own project; b) designing and developing new resources for the course, including audiovisual resources – this could

be partly done through Wikiversity using tools like Quiz extension or interactive approaches like those used in the Wikipedia Adventure ; c) designing and developing assignments, exercises and assessment procedures; d) adapting all educational resources to an open platform for MOOCs; delivering the course (about 6 weeks); e) evaluating the course performance (through participation data and a survey to all participants).

Task 5.5: Organization of training events (**Lead UM**; all partners; Month 1-36) Organization of training events on all aspects of Wikidata for research. These training events will be targeted, basically, at a) researchers – those based at partner institutions in the project, those participating in other EU-funded research projects, and those participating in citizen science projects, b) Wikimedians involved in Wikidata and other Wikimedia projects. Special attention will be paid to the subsets of both groups that are developers, as well as to their intersection, with the aim of expanding the ways in which Wikidata can be used, and nurturing communities of actual users. Some of the events will also be held at different conferences relevant to the project.

Deliverables

D5.1 Tutorials and Course Materials (Month 20)

D5.2 Massive Open Online Course (MOOC) on Wiki4R (Month 30)

D5.3 Training Events Report (Month 34)

D5.4 Dissemination and Stakeholder Engagement Report (Month 36)

List of work packages (Table 7)

Table 7.
List of work packages

Work package number	Work packagetitle	Lead participant number	Lead participant (short name)	Person-months	Start month	End month
WP1	Coordination and Management	1	MfN	13	1	36
WP2	Semantic modelling	2	UPM	25	1	36
WP3	Integrating research resources with Wikidata	3	UM	42	1	36
WP4	Enabling the use of Wikidata in research contexts	4	WMDE	60	1	36
WP5	Dissemination, stakeholder engagement and training	5	UOC	60	1	36
				200 Total months		

List of deliverables (Table 8)

Table 8.
List of deliverables

Deliverable (number)	Deliverable name	Work package number	Short name of lead participant	Type	Dissemination level	Delivery date
D1.1	Data Management Plan	WP 1	MfN	R	PU	M06
D1.2	Periodic report	WP 1	MfN	R	PU	M12
D1.3	Periodic report	WP 1	MfN	R	PU	M24
D1.4	Periodic and final report	WP 1	MfN	R	PU	M36
D2.1	Report on property profiles	WP 2	UM	R	PU	M24
D2.2	Report on semantic ontology mappings	WP 2	UPM	R	PU	M31
D3.1	Report describing quality assurance measures	WP 3	UPM	R	PU	M18
D3.2	Lessons learned from integration of TEL, BODR, and Recon2	WP 3	UM	R	PU	M22
D3.3	Report on other integration uses cases	WP 3	WMDE	R	PU	M33
D3.4	Report on the motivations and pitfalls around releasing open data	WP 3	MfN	R	PU	M35
D4.1	SPARQL endpoint for Wiki4R	WP4	UPM	OTHER	PU	M12
D4.2	Mechanism for semantic knowledge-base mirroring	WP4	UPM	OTHER	PU	M20
D4.3	Research-oriented Wiki4R edition of Wikidata software	WP4	UPS	OTHER	PU	M24
D4.4	Report on usability of Wiki4R in citizen science, digital humanities and cultural heritage research	WP4	MfN	R	PU	M33
D4.5	Report on Wiki4R integration and usability tests in science laboratories	WP4	UPS	R	PU	M36
D5.1	Tutorials and Course Materials	WP 5	UOC	R	PU	M20

D5.2	Massive Open Online Course (MOOC) on Wiki4R	WP 5	UOC	DEC	PU	M30
D5.3	Training Events Report	WP 5	UM	R	PU	M34
D5.4	Dissemination and Stakeholder Engagement Report	WP 5	MfN	R	PU	M36

Inter-relation of the main components of the project (Fig. 4)

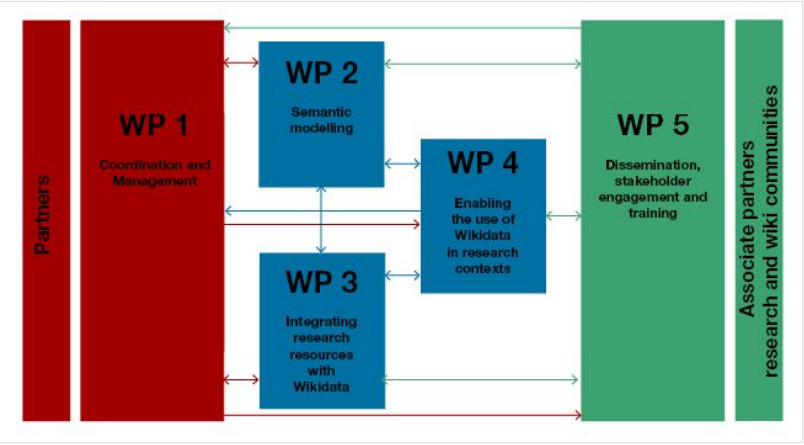

Figure 4.

Pert chart explaining the main interactions of the work packages.

Management structure and procedures

Organisational structure and decision-making

For implementing the project, the **Wiki4R** Consortium will initially comprise 7 partners from 4 EU member states. The project duration is planned as a period of three years (36 months).A project of this small complexity and scale requires an **appropriate management structure** and will be implemented by the project coordinator (MfN) within WP1 (Coordination and management). These structures as well as the **decision making processes** will be established according to the following principles:

- due consideration of the **equality** and **collective responsibility** of all participants,
- greatest possible **transparency** of the overall management, and decision making processes,
- ensuring **compliance** with all relevant provisions and regulations of the European Commission (EC),
- providing for **cost efficiency** and **professional administration**,

- realization of effective **sound monitoring** and **quality management**,
- undertaking research according to **best scientific practices**.

The proposed coordination structure for Wiki4R is based on long management experiences gained from other research projects and networks with similar requirements. The coordination and management structure for Wiki4R will be composed of the following core elements:

- General Assembly (GA)
- Steering Committee (SC)
- Project Coordinator (PC)
- Advisory Board (AB)

Their interrelationship and the overall Wiki4R management structures are illustrated by Fig. 5. The management bodies, their composition, responsibilities, scope of decision-taking power, and working procedures will be described in detail in the Consortium Agreement to be signed by all partners. It will be each partner's general responsibility to undertake all reasonable endeavours to perform and fulfil, promptly, actively, and on time, all of its obligations under the Grant Agreement and the Consortium Agreement as well as their tasks within the work packages according to their role in the Consortium including the submission of the deliverables as described in the proposal. The following structures and mechanisms will be essential elements for efficient management of the entire project.

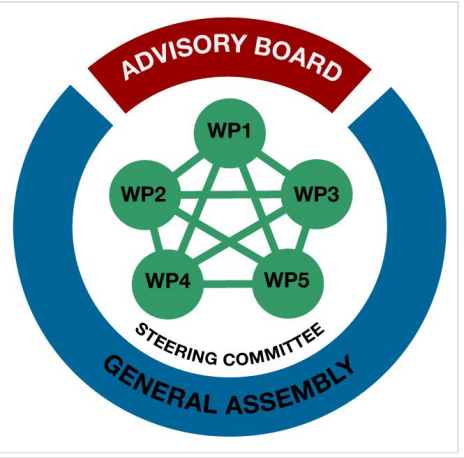

Figure 5.

Relations of Advisory Board, Steering Committee and General Assembly.

- due consideration of the **equality** and **collective responsibility** of all participants,
- greatest possible **transparency** of the overall management, and decision making processes,
- ensuring **compliance** with all relevant provisions and regulations of the European Commission (EC),

- providing for **cost efficiency** and **professional administration**,
- realization of effective **sound monitoring** and **quality management**,
- undertaking research according to **best scientific practices**.

The proposed coordination structure for Wiki4R is based on long management experiences gained from other research projects and networks with similar requirements. The coordination and management structure for Wiki4R will be composed of the following core elements:

- General Assembly (GA)
- Steering Committee (SC)
- Project Coordinator (PC)
- Advisory Board (AB)

Their interrelationship and the overall Wiki4R management structures are illustrated by Fig. 5. The management bodies, their composition, responsibilities, scope of decision-taking power, and working procedures will be described in detail in the Consortium Agreement to be signed by all partners. It will be each partner's general responsibility to undertake all reasonable endeavours to perform and fulfil, promptly, actively, and on time, all of its obligations under the Grant Agreement and the Consortium Agreement as well as their tasks within the work packages according to their role in the Consortium including the submission of the deliverables as described in the proposal. The following structures and mechanisms will be essential elements for efficient management of the entire project.

General Assembly

The entire **Wiki4R** partnership is represented in the General Assembly (GA), which is the **ultimate decision making body** for the consortium. The GA comprises all project partners, who are also eligible to vote on propositions presented to the GA. Each partner to the consortium sends **one official representative** and has **one vote**. The GA will be chaired by the Project Coordinator.

Decisions taken by the GA primarily generally relate to:

- budget-related matters,
- periodic activity reports and financial reports,
- periodic activity reports and financial reports,
- acceptance and addition of new partners,
- if required, exclusion of partners, updating the work plan, and modification of the Consortium Agreement.

If voting is required, decisions shall be taken by a majority of two-thirds (2/3) of the votes, unless otherwise provided in the Consortium Agreement. The GA can also appoint an **equity liaison person** who will be in charge of gender equity issues and will, if needed, take on the function of **ombudsperson** in case of conflict resolution between individuals.

Ordinary GA meetings will take place **once a year** in conjunction with major community events, such as Wikimania, Wikimedia Hackathons, or GLAM-Wiki meetings.

Steering Committee

The **Wiki4R** Steering Committee (SC) is the main **coordinating body** for the ongoing project execution and its performance in compliance with the work plan. The Steering Committee will work on behalf of the General Assembly, and regularly report to the GA on interim decisions taken and other items of interest. Ordinary **SC meetings** are to be held at least **twice a year**, and the SC will coordinate work inter-seasonally by conference calls or other means. Ordinary SC members are the Coordinator and the Work Package Leaders (WPL). The Coordinator shall chair all meetings of the SC, unless decided otherwise by the General Assembly. The Steering Committee will strive to work by consensus. In case of need of voting, decisions shall be taken by a **majority of two-thirds (2/3) of the votes**, unless otherwise provided in the Consortium Agreement. Each member of the SC will have **one vote**, and in case of equal votes the chair will have a casting vote. The Steering Committee is primarily responsible for supporting the Project Coordinator in fulfilling the legal and contractual **obligations** towards the European Commission by **monitoring and controlling** the efficient implementation of the project's objectives and work programme, on the basis of regular, 6-monthly financial and activity reports from each of the work packages. The SC will also provide for risk management and consider applying appropriate rules on "intervention measures" in case of significant delays or breach of obligations (details of this will be made explicit in the Consortium Agreement), and conflict management within the consortium.

Project Coordinator

The Project Coordinator (PC) is the **intermediary between the Wiki4R consortium** and the European Commission and is responsible for the **overall management** of the project, including the submission of reports and achievement of the deliverables. He ensures compliance by the partners with the contractual obligations in all activities of the project. The Project Coordinator has a right of veto in the case of decisions conflicting with the EC Grant Agreement at the General Assembly and at the Steering Committee. The coordinator will **administer the EC financial contribution** in accordance with the Grant Agreement, and the decisions taken by the consortium via the General Assembly.

The **Wiki4R** consortium will be led by the Museum für Naturkunde – Leibniz Institute for Evolution and Biodiversity Science (MfN). The project coordinator will be supported in this function by an **Administrative Project Manager**. This person will be in charge of day-to-day administration including the detailed **contractual and budget management, financial controlling, and reporting**.

The MfN is one of the World's leading natural history collections and research institutions and holds competence in areas of biodiversity and environmental research, from biogeography and taxonomy to information management, climate change, and advising

environmental policy. The museum has a staff of about 250 persons, an annual budget of about € 22M (2013), and it is currently undergoing a period of significant expansions and renovation. In 2009, the MfN was re-constituted as a Foundation under public law and became a member of the **Leibniz Association**, one of Germany´s four leading research organizations. The balance sheet and the accountancy of the MfN are checked annually by an independent **external auditor**.

External funding is handled by the Third Party Funds Department, which manages more than 100 externally funded projects, among them several EC-funded projects (e.g., coordinating EU BON (Co. 308454), and as Partner in SYNTHESYS 3 (Co. 312253), OpenUp! (Co. 270890), European Creative (Co.325120), pro-iBiosphere (Co. 312848) and other FP7 finished projects.

The MfN has its own **Bureau of Media Communication**, which maintains regular contact with different media, politicians, and scientific organizations, and will be a valuable support mechanism for public outreach, and the broader dissemination of the project´s results. Finally, the Project Coordinator will also be supported by the **EU Liaison Office of the Leibniz-Association in Brussels**. The office will advise the Coordinator and his team in any EU policy related issues as well as in relevant administrative questions concerning Horizon2020.

Work Package Leaders

Each Work Package will have one designated Work Package Leader (WPL), who is taking on this role in addition to his/her role as an individual partner to the project. WP leaders report to the Project Coordinator. They are appointed by the institutions designated to lead the WP have sound scientific background and wide management experience to guide the respective WP and tasks, and to ensure performance and progress of the work with regard to project deliverables and milestones. In particular, WP leaders will be responsible for:

- effective **communication between members** of the WP and to the Steering Committee,
- for effective (within designated resources) and timely (within designated deadlines) completion of work, milestones, and deliverables
- any activities, deliverables and information as relevant to other partners or the consortium as a whole;
- **quality control and technical management**, i.e. the coordination and monitoring on a day-to-day basis of progress of the individual work package and its task, including;
- delivering **interim activity reports** i.e. referring to work progress and budget-development, and alerting in case of delay or default and if necessary with suggestions for the solution of possible challenges and problems.

The following WPL have been assigned and agreed: Table 9.

Table 9.
Work package leaders

Work package No.	Work Package description	WP Leader and member of the Steering Committee	Beneficiary Acronym
WP1	Coordination and management	Dr. Gregor Hagedorn	MfN
WP2	Semantic modelling	Prof. Dr. Asunción Gómez-Pérez	UPM
WP3	Integrating research resources with Wikidata	Dr Egon Willighagen	UM
WP4	Enabling the use of Wikidata in research contexts	Lydia Pintscher	WMDE
WP5	Dissemination, stakeholder engagement and training	Dr. Eduard Aibar	UOC

Advisory Board

The main role of the Advisory Board (AB) will be to help to guarantee the delivery of high scientific and technical quality output from the project, and to ensure that the project partners will not become isolated within their own thinking and perceptions. In that function, the AB will act as a constant internal review mechanism, advising independently with both quality control and risk management. While generally being asked on views and advice on specific issues, the AB will be free to express its opinion on any aspect of the project's activity and performance.

The AB will also advise on the outreach to the wider scientific and stakeholder communities not directly involved in the project and for benefiting from independent, external expertise. The AB will be established at the beginning of the project.

The AB may call on additional experts on an ad hoc basis, should the need arise. For the full duration of the project, members of the AB are expected to be available who can be contacted at any time for specific questions. The AB will primarily advise and work with the Project Coordinator and Steering Committee, but deliver its main reports to the General Assembly. For its work, the AB will also liaise with the internal project working groups.

Consortium Agreement and IPR management

A Consortium Agreement (CA) will be developed and signed by all partners, which will set out in detail the rights, responsibilities, and liabilities of participants to each other and to the consortium. In addition to detailed provisions for consortium organization and governance, it will include recommendations for equality and gender issues, Intellectual Property Rights (IPR), data exchange, use of software and other resources within the project, promoting open access and free sharing of information. It will be based on the DESCA model for Horizon2020 projects, and include relevant provisions to also ensure quality control. Conflict Management resolutions will also be addressed in detail in the CA.

Communication within the Consortium and Project Meetings

Active communication process will be installed to ensure internal coherence, and effectiveness in the daily exchange of information and cooperation for all consortium partners. It will be the responsibility of the Coordinator to facilitate efficient communication for the project, which will rely on the following mechanisms:

- email correspondence as the primary means of day-to-day communication, including use of mailing lists and electronic discussion forums for the project and individual communities,
- regular conferences, using internet-based video conferencing tools,
- communication through the project web- platform.

In addition frequent and regular meetings at various levels are regarded as key to the success of projects such as Wikidata for research, and for developing a good and lasting network of networks. The following Table 10 lists relevant meetings to take place during the course of the project.

Table 10.
List of project meetings

Persons/Bodies	Meeting frequency	Objectives of the Meeting
SC	monthly (online)	operational decisions, monitoring progress
GA	yearly (on-site, in conjunction with relevant community meetings)	reviewing progress, fundamental decisions, community engagement
AB	back-to-back with GA	review mechanism, advice on further developments

The list of milestones is available from Table 11.

Table 11.
List of milestones

Milestone number	Milestone name	Related work package(s)	Estimated date	Means of verification
MS1.1	Kick-off meeting	WP1	M02	Meeting held, minutes of the meeting
MS1.2	Project web platform established	WP1	M02	Up and running
MS1.3	Project meeting as satellite to Wikimania conference	WP1	M14	Meeting held, minutes of the meeting
MS1.4	Project meeting as satellite to Wikimedia Hackathon	WP1	M30	Meeting held, minutes of the meeting
MS1.5	Review meeting	WP1	M14	Review held
MS1.6	Review meeting	WP1	M26	Review held

MS2.1	Profile and mapping for the first use case (TEL)	WP2	M04	Inclusion in Wikidata newsletter
MS2.2	Profile and mapping for the second use case (BODR)	WP2	M08	Inclusion in Wikidata newsletter
MS2.3	Profile and mapping for the third use case (Recon2)	WP2	M12	Inclusion in Wikidata newsletter
MS2.4	Profile and mapping for the remaining use cases in **T3.1**	WP2	M20	Inclusion in Wikidata newsletter
MS3.1	Data citation mechanism established	WP3	M06	Data citation demonstrated
MS3.2	Survey of potential further use cases for Wiki4R completed	WP3	M08	Results presented
MS3.3	Error diagnosis mechanism established	WP3	M09	Error diagnosis demonstrated
MS3.4	Integration of TEL use case	WP3	M09	Integration demonstrated
MS3.5	Quality review mechanisms established	WP3	M12	Review and verification demonstrated
MS3.6	Survey of motivations for further use cases completed	WP3	M12	Integration demonstrated
MS3.7	Decision which further use cases to select for Wiki4R	WP3	M14	Minutes of SC meeting
MS3.8	Integration of BODR use case	WP3	M15	Integration demonstrated
MS3.9	Integration of Recon2 use case	WP3	M21	Integration demonstrated
MS4.1	Survey of use cases for Wikidata within relevant research communities and cultural heritage institutions	WP4	M6	Publicly available
MS4.2	Wikidata game to strengthen citizen engagement available	WP4	M20	Publicly available
MS4.3	Packaging of mirrored Wiki4R in virtual machine M24	WP4	M24	Publicly available
MS4.4	Prototype of unit tests for data established	WP4	M30	Publicly available
MS5.1	Production of first set of tutorials and course materials	WP5	M10	Tutorials and course materials online
MS5.2	Test version of MOOC available	WP5	M16	Testable MOOC environment
MS5.3	MOOC delivered	WP5	M35	Number of students enrolled; students satisfaction survey

The management of risks is detailed in Table 12.

Table 12.
Management of risks

Description of risk	Involved WP(s)	Proba-bility	Im-pact	Proposed risk-mitigation measures
Slow recruiting procedures at beneficiary level	Any	High	High	Start recruiting preparations before the project has started
Bankruptcy or withdrawal of partner	Any	Low	High	Initiate discussions with EC to define a solution allowing the implementation of the work plan. Delegate work to existing partners or new subcontractors or external partners.
Partner is unresponsive	Any	Low	High	Delegate work to other partners with expertise and capacity. Obtain a SC resolution to transfer resources away from unresponsive partner.
Unattainable Deliverable	Any	Low	High	Investigate alternative options in the context of the high-level objectives of Wiki4R and the expertise and capacity of partners. Discuss with the EC PO.
Delayed Deliverable	Any	Medium	Medium	Investigate causes for delay, increase monitoring intensity by the respective SC and the WPL, document causes and discuss with EC PO.
Delayed Milestone	Any	High	Low	Assess dependencies and prioritise all dependent tasks to avoid a domino effect, assess options to change the order of work.
Lack of consistent ontologies to support the semantic mapping	WP	Low	Low	The system will be able to work with semantics which only exist in Wikidata and which have no external equivalents.
Proposed class/ property profiles conflict with citizen community plans	WP2	High	Medium	Work with the community from the start; be pragmatic; record problems and report them publicly. Last resort is the primary use of Wiki4R mirror instances in these use cases.
Datasets proposed for inclusion in Wikidata conflict with community plans	WP3	High	High	Engaging with the community, especially the relevant WikiProjects, with transparent and responsive communication. Last resort is the primary use of Wiki4R mirror instances in these use cases.
The workflow integration does not satisfy the researchers	WP4	High	Medium	With an open and agile development model, we will be able to respond to unforeseen demands. A reallocation and reprioritization of resource may be necessary. The impact is medium because it does not create a dependency in the project, and the need of continuous work on Wiki4R beyond the project duration is already anticipated.

Infrastructure unsustainable beyond end of project	All	Low	High	Most of the critical infrastructure is guaranteed by the Wikimedia Foundation with a separate and tested sustainability model. Infrastructure can become unsustainable if computing intensity exceed reasonable resources for investments into computing infrastructure; in this case software strategies must be improved or the use cases limited.
Initial near real-time synchronization is relatively hard to achieve, since Wikidata is edited about 2.5 times per second	WP4	High	Low	This is a hard-to-assess problem depending on many factors. Its immediate risk on the project is low because it does not create a dependency: The first implementation can have a delay of 10 minutes between the modification through a scientist and the reuse of this modification in her tools.
Near real time synchronization after optimization is still too slow to fulfil use case requirement of use cases	WP4	Low	Medium	For full establishment and acceptance, a low near-real time on the order of few seconds will be required. Technical solutions are known but remain to be robustly implemented and tested for the specific workload. Should technical solutions fail, the use cases need to be analysed for priority demand with the goal to achieve a faster synchronization for a subset of information items.
Failure to engage sufficient numbers of scientists in interacting with the VRE	WP	Low	High	Test the relative efficiency of the various engagement and communication strategies to reallocate resources to the more effective ones.
Low engagement from the research community and low use of training materials	WP5	Low	High	Targeting training sessions and dissemination events to particular research communities and developing modular training materials able to be adapted for different audiences.
Work planned by Wiki4R is carried out by someone else	All WPs	Medium	Low	Since we are working in the open, this is not unexpected, and we would welcome it. Reallocate resources in view of overarching Wiki4R goals.
Wiki4R is not adopted by the community in the expected time frame	WP4, WP5	Low	High	Wiki4R, addressing both professional and citizen scientists, is probably unique in that a huge community in terms of citizen scientists is already there, i.e. citizens volunteering their spare time to curate data. We cannot guarantee yet a huge community of professional scientists, as we are focussing on pilots. For these pilots, however, the participation is already secured. While not endangering the Wiki4R project itself, lack of professional adoption beyond pilots would destroy its impact. The work plan is therefore carefully planned with this in mind and significant resources are invested in training and capacity building to make this less likely.

Consortium as a whole

The consortium brings together seven partners whose core competencies span the major branches of the natural sciences and into the information sciences as well as the cultural and natural heritage sector, and civil society.

With the exception of WP1, all partners are involved in all work packages, which ensures effective interaction beyond meetings or formal communications.

The profiles of the partners are distinct, but overlap in a systematic fashion:

- UPM, WMDE, UM and UPS are all providers of semantic technologies, and while UPM and WMDE work across disciplines, they are distinguished by a focus on automated techniques (UPM) and community-supporting platforms (WMDE). Both are highly involved across WP2-4.
- Similarly, while both UM and UPS work in chemistry, the former bridges towards computational chemistry and biology, the latter towards experimental chemistry and the physical sciences. UM focuses on the integration of information into Wikidata (WP3), UPS on the reuse of information from Wikidata (WP4).
- UPM, UM and UPS are campus universities, whereas the fourth university in the consortium, UOC, is a virtual university. UOC will thus focus on the virtual training aspects, with UM and UPS contributing. All partners except UOC will organize on-site training events, and all partners will contribute to the creation of course materials as well as to dissemination and community engagement (WP5).
- UOC has in the past performed research on the use of Wikipedia in higher education. Its participation in the other work packages thus will not only inform the training activities in WP5 but also a new line of research on the use of Wikidata in higher education.
- EF is a data provider and as such involved in WP3, but its focus within the project is on exploring the potential of using Wikidata in cultural heritage contexts (WP4), together with MfN; this builds on prior collaboration in former EU projects like OpenUp!
- Beyond its coordinator role, MfN provides multiple aspect of biological expertise. It is involved in all work packages, as a data provider (WP3), data re-user (WP4) and use case provider for the semantic modelling (WP2). Use cases are from biology (e.g. taxa), mineralogy, as well as cultural history (historic records). Citizen science is transdisciplinary. Finally, it addresses sustainability issues in “Optimizing openness” (T3.3), which builds the foundation for future projects of this kind.

Most but not all of the key personnel have contributed to Wikidata in the past, and all are experienced in EU research projects.

The partners are complemented by 17 associate partners, as detailed in section 4.3 of the proposal.

Most of them are based within the European Research Area, but some (Scripps, MetaBrainz and Center for Open Science) in the US. As a group, their expertise goes well beyond that of the consortium and covers areas ranging from genetics to music technology, from history to mathematics, from data mining to publishing, libraries and repositories. They also cover the value chain from basic research to applied research to scholarly societies and professional bodies, from non-profit to startup, from botanic garden to universities.

They will provide use cases or data, advice on semantic modelling, data integration or data reuse within their fields, their countries or for their languages, they will help engage their respective communities, and they will contribute to training and dissemination activities. Importantly, the associate partners will also serve as seeds for follow-ups to Wiki4R: our proposal is openly licensed and intended to be forked, so as to help build an ecosystem of Wiki4R projects and initiatives that are linked on the one hand to Wikidata either directly or through the infrastructure we will build, and on the other hand to their respective communities.

Resources to be committed

In consideration of the size of the project, careful attention has been given to a realistic allocation of the Wiki4R budget among partners, reflecting their respective involvement and work load in each of the WPs. The budget as presented here has been agreed by the consortium and all partners.

The total costs of the project amount to 1,553,685.00 Mio €, with an EC-contribution of 1,553,685.00 Mio €. The project will have a duration of 3 years (36 months) and requires 200 person months in total.

The total costs as given in Table 13 relate to the auditable aggregated costs that will be incurred by all partners, corresponding to the administrative forms.

Table 13.
Table of resources

	Participant							**Total**
	MfN	**UPM**	**UM**	**WMDE**	**UOC**	**EF**	**UPS**	
	Country							
	DE	ES	NL	DE	ES	NL	FR	
Direct personnel costs / €	280,200.00	100,000.00	223,200.00	180,828.00	106,700.00	136,620.00	120,000.00	1,147,548,00

Other direct costs / € + costs for in-kind contribution not used on the beneficiary's premises	35,000.00	8,000.00	17,000.00	8,000.00	7,000.00	11,000.00	7,000.00	93,000.00
Direct costs of sub-contracting / €	3,000.00	0	0	0	0	0	0	3,000.00
Costs of in-kind contributions not used on the beneficiary's premises / €	0	0	0	0	0	0	0	0
Indirect Costs / € (=0.25 (A +B-E))	78,800.00	27,000.00	60,050.00	47,207.00	28,425.00	36,905.00	31,750.00	310,137.00
Special unit costs covering direct & indirect costs	0	0	0	0	0	0	0	0
Total estimated eligible costs / € (=A + B+ C+ D+ F+ G)	397,000.00	135,000.00	300,250.00	236,035.00	142,125.00	184,525.00	158,750.00	1,553,685.00
Reim-burse-mentrate	100%	100%	100%	100%	100%	100%	100%	
Max. Grant / € (=H*I)	397,000.00	135,000.00	300,250.00	236,035.00	142,125.00	184,525.00	158,750.00	1,553,685.00

Requested grant / €	397,000.00	135,000.00	300,250.00	236,035.00	142,125.00	184,525.00	158,750.00	1,553,685.00

A summary of staff effort is given in Table 14.

Table 14.
Summary of staff effort

	WP1	WP2	WP3	WP4	WP5	Total Person/ Months per Participant
Participant 1/ MfN	**9**	3	12	13	11	48
Participant 2/ UPM	1	**6**	5	4	4	20
Participant 3/ UM	1	6	**13**	4	12	36
Participant 4/ WMDE	1	6	7	**14**	8	36
Participant 5/ UOC	1	1	1	2	**15**	20
Participant 6/ EF	0	2	3	10	5	20
Participant 7/ UPS	0	1	1	13	5	20
Total Person/Months	13	25	42	60	60	200

Human Resources

Personnel costs represent the largest cost category for Wikidata for research, reflecting the labour intensive RTD nature of the project. For Wikidata for research, mainly postdoctoral researchers and technical staff will be required and offered positions, mainly to be recruited from universities and research institutions involved. For the management tasks, the coordination team will be supported by a 25% project management position to be paid by the project. In addition, the project will benefit greatly from considerable work time and support provided in kind by personnel, predominately senior staff of professorial or equivalent status, from many partner institutions, including several of the WP leads. Their combined effort and commitment in all areas of the project will add a substantial number of additional person months and resources, not accounted for in the declared budget.

Equipment and Consumables

Most Wiki4R project partners can and will rely on their existing institutional ICT infrastructure and equipment for the project, less than 1% budget has been allocated to equipment. Additional equipment needed means primarily ICT hardware and software, and related tools and working stations.

A small amount of the requested budget will be used for consumables. Consumables relate to the preparation of training materials and tutorials planned under WP5. 2,000€ for Partner have been planned to cover open access fees. In addition to that, the management budget

of the Coordination in WP1 includes an open access fees contingency fund (ca 8,000€) available to all partners for project-related publications.

Travel costs

Travel costs include mainly attendance of project **meetings** for the General Assembly, Steering Committee as well as other for groups and committees as foreseen in the work plan. WP-Leaders have a budget of 6,000€, whilst the other partners 5,000€ foreseen for the three years.

The management budget of the Coordination (WP1) includes a travel contingency fund to allow to cover travel expenses of members or representatives of the Advisory Board and of the associate partners, as well as for invitation of additional stakeholders to project meetings.

Sub-contracting

All **sub-contracting** will follow Horizon 2020 financial guidelines and is related to rules for awarding contracts according to the principles of best value for money (best price-quality ratio), transparency, and equal treatment.

Sub-contracting costs refer to those partners with an EC contribution higher than 325k€ that will be subject of financial audits following the financial rules under Horizon 2020. These audit costs are planned in WP1, under sub-contracting and vary between 2000 and 3000€ per audit for MfN.

Members of the consortium

Participants

Partner 1 (Coordination): Museum für Naturkunde (MfN), Germany

The Museum für Naturkunde Berlin (MfN) – Leibniz Institute for Evolution and Biodiversity Science is a research museum within the Leibniz Association and constituted as a foundation under public law. It is one of the most significant research museums worldwide focusing on biodiversity, evolution and geo-sciences with over 250 staff members. Research at the Museum für Naturkunde is organized in four Science Programs: Evolution and Geoprocesses, Collection Development and Biodiversity Discovery, Digital World and Information Science, and Public Engagement with Science. The collections of the MfN are directly linked to research and comprise more than 30 million specimens relating to zoology, palaeontology, geology and mineralogy. In addition, the MfN houses a unique Animal Sound Archive containing approximately 120,000 animal sound recordings. The library of the Museum für Naturkunde Berlin is one of the most important reference libraries in zoology in the German-speaking world.

Relevant experience and role in project

MfN has been participating and taking a leading role in several EU-projects such as EU BON, SYNTHESYS III (I3), the European Distributed Institute of Taxonomy (EDIT), European Biodiversity Heritage Library (BHL), OpenUp!, 4D4Life, ViBRANT, pro iBiosphere and Europeana Creative, as well as other international collaborative research programmes (e.g. BIOTA). It hosts the secretariat of the European Citizen Science Association (ECSA) and has extensive experience in citizen science. It is also a member of the Science 2.0 research network of the Leibniz Society. MfN coordinates the Wikidata for research project and provides for professional administration, management of the consortium, as well as targeted implementation of the project goals (WP1). It will also take a leading role in the assessment of the benefits of and barriers to openness (WP3) as well as the tasks related to citizen science and cultural heritage (WP4) as well as in dissemination, communication (WP5).

Profile of the personnel

Dr. Gregor Hagedorn (m) is the head of the research division “Digital World and Information Science” at the MfN. He has extensive experience in developing data standards (three Biodiversity Information Standards/TDWG.org standards), descriptive and trait data, computer aided identification and citizen involvement (German Open Nature Guides, Artenquiz). Current projects are “German Federation for the Curation of Biological Data” (GFBio), pro-iBiosphere (EU Project) and Europeana Creative. He is a member of the CETAF and GBIF-Germany IT-commissions.

Dr. Daniel Mietchen (m) is a biophysicist specialized in non-invasive imaging, with complementary research interests ranging from evolutionary aspects of vocal production to semantic integration of biodiversity literature. He has broad experience in disseminating scientific information within and beyond the scientific community through open licenses, e.g. by way of data publishing or reuse in educational contexts like Wikimedia platforms. From 2011-2013, he served as Wikimedian in Residence on Open Science at the Open Knowledge Foundation Germany. As a volunteer, he leads a team operating a software that collects openly licensed audio and video materials from the biomedical literature repository PubMed Central and uploads them to Wikimedia Commons for reuse on Wikimedia platforms. The team was one of the inaugural recipients of the Accelerating Science Awards, and that work has recently been expanded to a prototype for full-text upload to Wikisource.

Falko Glöckler’s (m) expertise is biodiversity informatics, especially data workflows, databases, web developments and biodiversity data standards. He is in charge of data integration in several digitization projects and standardization and mobilization of collection and field data for providing these to international aggregation platforms (GBIF, Europeana). His specialities are data quality management and compliance algorithms for biodiversity data.

Dr. Hatem Mousselly Sergieh (m) obtained 2014 a double PhD degree in computer science from the University of Passau – Germany and INSA Lyon – France. He has expertise in data mining, ontology matching and information retrieval and works as data scientist.

Ralf Thomas Schmitt (m) is curator for minerals and rocks at the Museum für Naturkunde Berlin since 1994. He is responsible for digitization and the continuous development of the mineralogical database. As part of a SYNTHESYS networking activity, he was involved in the development of the ABCDEFG (Access to Biological Collection Databases Extended for Geosciences) XML-Schema for standardization of mineralogical, geological and palaeontological digitized collection data. This scheme is successfully used for the online publication of geoscientific collection data from the Museum für Naturkunde Berlin via the GEOCASE portal, which is combining collection data of several geoscientific institutions from Europe.

Publications

1. Hagedorn G, Mietchen D, Morris R, Agosti D, Penev L, Berendsohn W, Hobern D 2011. "Creative Commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information". ZooKeys 150 (150): 127–149. DOI: 10.3897/zookeys.150.2189.
2. Egloff W, Patterson D, Agosti D, Hagedorn G 2014. Open exchange of scientific knowledge and European copyright: The case of biodiversity information. ZooKeys 414: 109135. DOI: 10.3897/zookeys.414.7717.
3. Mietchen D 2014. "The Transformative Nature of Transparency in Research Funding". PLoS Biology 12 (12): e1002027. DOI: 10.1371/journal.pbio.1002027.
4. Penev L, Hagedorn G, Mietchen D, Georgiev T, Stoev P, Sautter G, Agosti D, Plank A, Balke M 2011. "Interlinking journal and wiki publications through joint citation: Working examples from ZooKeys and Plazi on Species-ID". ZooKeys 90. DOI: 10.3897/zookeys.90.1369.
5. Glöckler F, Hoffmann J, Theeten F 2013. The BioCASe Monitor Service - A tool for monitoring progress and quality of data provision through distributed data networks. Biodiversity Data Journal, 1: e968. DOI: 10.3897/BDJ.1.e968.

Relevant previous projects or activities

1. EU BON (FP7), Building the European Biodiversity Observation Network
2. Open-Up (EU ICT-PSP), Opening up the Natural History Heritage for Europeana
3. GBIF D, Global Biodiversity Information Facility – Germany
4. EU FP7 4D4Life, Distributed Dynamic Diversity Databases for Life
5. ECSA - European Citizen Science Association

Partner 2: Universidad Politécnica de Madrid (UPM), Spain

The **Universidad Politécnica de Madrid (UPM)** is the largest Spanish technological university. With two recognitions as Campus of International Excellence, it is outstanding in its research activity together with its training of highly-qualified professionals, competitive at an international level. More than 2,400 researchers carry out their activity at the UPM, grouped in 216 Research Groups, 10 Research Centers and 55 Laboratories, all of them committed to transforming the knowledge generated into advances applied to the production sector.

The intense collaboration with governmental bodies and industry guarantees that research at the UPM offers real solutions to real-world problems. The dynamism of research and development/innovation activity at UPM, together with the transfer of knowledge to society, is among its lines of strategy. These two commitments place it among the Spanish universities with the greatest research activity and first in the capture of external resources in a competitive regime. UPM heads the Spanish Universities' participation in the 7th European Framework Program with more than 280 projects and more than €80M funding. Moreover, every year, UPM applies for around 40 patents and receives a similar number of concessions demonstrating a high commitment to innovation. UPM is leader in business creation, having generated around 140 businesses. Its support and backing of the business sector is very close and it annually signs around 600 contracts with private businesses.

UPM is an institution **committed to the transfer of knowledge generated through its research structures to society**, and **its transformation into advances and technological developments applied to the productive sector.**

Relevant experience and role in project

The **Ontology Engineering Group (OEG)**, led by Prof. Dr. Gómez-Pérez, has been working on the provision of **ontologies, semantic infrastructures, multilingualism, sensor data and linked data** since 1995. It is composed of 30 researchers with a consolidated reputation in the fields of: Ontological Engineering, Data Integration, Linked Open Data, Semantic Web, NLP and Semantic e-Science. The OEG has coordinated seven European projects. Members of the OEG participated in more than 15 EU IST projects. OEG members participate on more than 15 working groups at the Consortium of the World Wide Web. The OEG has participated in 19 research and technology transfer contracts with Spanish and International companies and organizations (e.g. AENOR, OMS, FAO, Fujitsu, Telefonica, Indra, Atos Origin, National Spanish Library) and has collaborated with relevant international research groups in prestigious universities worldwide. The OEG hosts the Spanish chapter of DBpedia.

The group's contributions to the project will leverage its expertise in Semantic Web and Linked Data technologies and focus on the property profiles and semantic mapping (WP2), multilingual matters as well as quality assurance (WP3) and providing Wikidata as Linked Open Data (WP4).

Profile of the personnel

Prof. Dr. Asunción Gómez-Pérez (f) is Full Professor at UPM, director of the Artificial Intelligence department, director of the OEG and PhD in Computer Science (1993). Before joining UPM, she was visiting (1994-1995) the Knowledge Systems Laboratory at Stanford University. She also was the Executive Director (1995-1998) of the AI Laboratory at the School of Computer Science. She has coordinated SEALS, SemSorGrid4Env and Ontogrid and now she is coordinating LIDER. She has participated in more than 15 EU (Admire, Esperonto, etc.) and Spanish R&D projects (CENITS mIO!, España Virtual, Buscamedia, myBigData, GeoBuddies). Her main research interests are **ontologies, Linked Data, Multilingual Linked Data and the Semantic Web**. She has published more than 150 papers and two books on Ontological Engineering. Her works on Ontological Engineering about Methontology and the NeON methodology are known worldwide. Her works on Ontological Engineering about Methontology and the NeON methodology are worldwide known. She is primarily responsible for UPM task.

Dr. Jorge Gracia del Rio (m) is a postdoctoral researcher at the Artificial Intelligence Department at Universidad Politécnica de Madrid (UPM). He obtained his degree in Physical Science and his PhD in Computer Science at University of Zaragoza. His research experience has focused on **semantic measures, ontology matching, and disambiguation techniques** in the field of the Semantic Web. He has been research visitor at Knowledge Media Institute (Open University, United Kingdom) and INRIA (France). He has worked in several European projects: Dynalearn, Monnet, and currently in LIDER (where he develops the role of Quality Assurance Coordinator). His current research topic is **multilingualism on the Web of Data**. He co-chairs the W3C Best Practises on Multilingual Linked Open Data (BPMLOD) community group, and collaborates actively in the W3C Ontology Lexica (ONTOLEX) and Linked Data for Languages Technologies (LD4LT) community groups. You can find out more about Jorge Gracia on his website. He is primarily responsible for carrying out the proposed research related to multilinguality.

Dr. Mariano Rico (m) is a physicist (master degree) and computer scientist (PhD) with 10 years of experience in private IT companies and 10 years of academic experience. He is member of the DBpedia Internationalization committee and the person responsible for the Spanish chapter of DBpedia. He has organized three local workshops to disseminate the technologies related to the Semantic Web and Linked data in the context of the Spanish DBpedia. He has been researching in collaborative (wiki) technologies, virtual environments and the relation between linked data and multilingual natural language processing. He is primarily responsible for carrying out the proposed tools related to linked data generation.

Ms.C. Nandana Mihindukulasooriya (m) is a Software Engineer who is working for the Center for Open Middleware and the Ontology Engineering Group, Universidad Politécnica de Madrid. He is a member of the Ontology Engineering Group since 2011 and he has participated in the SEALS FP7 project and in the ALM iStack project. His main interests are

Linked Data, RESTful design, application integration and transaction processing. He is a member of W3C Linked Data Platform (LDP) WG and he is the editor of the LDP Primer. He was awarded a prestigious two-year Erasmus Mundus scholarship from the European Union during his double-degree M.Sc. from BTH (Sweden) and UPM (Spain). Nandana successfully completed three Google Summer of Code (GSoC) projects in 2006, 2007, and 2010. Previously, he worked as a Technical Lead at WSO2 Inc. where he played a leading role in the design and implementation of WSO2 Web Services and Cloud middleware platforms for which he was awarded the “Outstanding Contributor of the Year Award” in 2008. He was also a member of Web Services Secure Exchange (WS-SX) and Web Services Federation (WS-FED) technical committees in OASIS during that time. Nandana is an active contributor to open source and currently serves as a committer, Project Management Committee member and a mentor for several Web Services and Semantic Web related projects in Apache Software Foundation and was a speaker at ApacheCon USA 2009 and ApacheCon Europe 2012. He is primarily responsible for carrying out the proposed tools related to LDP.

Publications

1. Gómez-Pérez A., Fernández-López M, Corcho O (2003) Ontological Engineering. Springer-Verlag. November 2003.
2. Fernández-López M, Gómez-Pérez A, Suárez-Figueroa MC (2013) Methodological guidelines for reusing general ontologies. Data & Knowledge Engineering. Editorial Elsevier. ISSN: 0169-023X. July 2013.
3. Gracia J, Montiel-Ponsoda E, Cimiano P, Gómez-Pérez A, Buitelaar P, McCrae J (2012) Challenges for the Multilingual Web of Data. Journal of Web Semantics, 11, pp. 63-71. ISSN 1570-8268
4. Gracia J, Mena E (2011) Semantic Heterogeneity Issues on the Web. IEEE Internet Computing, ISSN 1089-7801, DOI 10.1109/MIC.2011.129, volume 16, number 5, pp. 60-67, September-October 2012. Available online at IEEE Xplore digital library.
5. Rico M, Gómez-Perez A (ongoing) The Spanish chapter of the DBpedia. Text version: http://es.dbpedia.org. SPARQL endpoint: http://es.dbpedia.rg/sparql. Dataset information: http://datahub.io/dataset/dbpedia-es

Relevant previous projects or activities

1. **http://es.dbpedia.org/** The OEG-UPM serves world-wide the Spanish Chapter of DBpedia, which is relevant for the project. This server stores 200 million RDF triples that received 22 million SPARQL queries in year 2013, and the web site had around 6000 different visitors (8000 visits), 20% with non-Spanish browsers.
2. Members of the OEG have participated at the **W3C Linked Data Platform Working Group and have developed LDP4j** (http://www.ldp4j.org/). LDP aims at describing a set of best practices and simple approach for a read-write Linked Data architecture, based on HTTP access to web resources that describe their state using the RDF data model. OEG members are active in the working group and contribute as editors of the LDP Test Cases, LDP Primer, and LDP Best Practices and Guidelines.

3. **Datos.bne.es** (http://datos.bne.es): This linked data project started by the end of 2011 has already published 10 million bibliographic and authority records as Linked Open Data under a Public Domain license (Creative Commons CC0), generating more than 60 million triples and more that 1 million links. Furthermore, it provides a portal provides full-text, entity-oriented and faceted search and navigation for end-users using the latest web technologies and with a focus on user experience.
4. **LIDER** (Project coordinated by UPM – A. Gómez-Pérez - Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe FP7-610782, support action). LIDER's goal is to define best practices and a reference architecture, develop a roadmap, as well to create a community that can support the release of **language resources as Linguistic Linked Open Data**, to foster the discoverability of such datasets as well as their exploitation by linked-data-aware natural language processing and content analytics services. LIDER is seeking to network with projects in the area of Natural Language Processing, Linked Data and Big Data Analytics. Years: 2013-2015, Value: €693,775
5. The previous most connected project to the subject of this proposal in which UPM was involved is the **NeOn project** (Lifecycle support for networked ontologies - FP6-027595) aimed to advance the state of the art in Ontology Engineering and Semantic Web technologies. The main goal was to provide effective methodological and tool support for developing a new breed of semantic applications, able to exploit effectively the large amounts of information and data, which are now available on the web. Years, 2006-2010, Value: €1,135,822.
6. **Monnet** (Multilingual Ontologies for Networked Knowledge FP7-248458). This project provided a **semantics-based solution for integrated information access across language barriers**. As outcome, Monnet developed semi-automatic approaches and methodologies allowing cost effective ontology localization as a basis for implementing an integrated solution to providing semantic-level access to information across languages. One of the results was the lemon model for modelling lexicon and machine-readable dictionaries, linked to the Semantic Web and the Linked Data cloud. Years: 2010-2013, Value: €362,456

Significant infrastructure

Thanks to the extensive trajectory in developing and managing R&D projects, the Ontology Engineering Group may provide a wide set of tools and methods that may benefit the project from a scientific and technical point of view.

- A system administrators' team supports the project technical team in acquisition, configuration and installation of required hardware equipment, IT services or software licenses.
- A professional project management office supports project managers in the financial reporting and auditing processes that take place during the project.
- A set of collaborative tools is at the disposal of project team for communication (online meeting tool, mailing lists manager), for management (budget/payments monitoring spreadsheets, project plan and quality control methods), for IPR

management (IPR control spreadsheet) or for software development (software configuration management, change management, continuous integration and delivery, quality control).

- In-house virtualized servers hosting software that support a modern software development lifecycle (SDLC), including but not limited to: Software configuration management (Git, Mercurial, and Subversion); Software project management and collaboration system (Redmine); Software delivery (Nexus); Various state-of-the-art tools and the related expertise for implementing automated testing (unit, integration, functional, stress/stability/scalability); VMs for hosting complex development and testing environments
- Hosted (SaaS) services related to the SDLC, including but not limited to: Software configuration management and related collaboration services (simple wiki, simple issue tracking, social collaboration features) like GitHub and BitBucket; Modern software development collaboration environment based on Jira and Confluence; Virtual machines used for testing software systems
- Software libraries and modules that can provide the infrastructure for implementing more complex software artefacts and applications for the proposed project.
- A technical library of paper and electronic books related to technical areas relevant to the proposed work.

The OEG infrastructure in a nutshell:

- 19 servers and 8 PCs acting as a server.
- SEALS infrastructure: 7 servers and 1 SAN (storage). 3 servers for virtualization using VMware vSphere 4.1, and 4 servers with different roles (1 frontend server and 3 backend servers). Roles: DHCP, DNS, RRAS, DFS, fibre channel nodes, terminal services, balancer, preproduction/beta test/final services.
- Linked Data infrastructure: 4 servers. SPARQL endpoints with Virtuoso, pubby and some CMS (Joomla, Drupal).
- OEG dedicated: 7 servers with internal/external services (email with Postfix, Mediawikis, Jira, Bamboo, FishEye & Crucible, dedicated virtualization server, FTP, databases and file server. Other server/PCs are fully dedicated to other projects or initiatives.
- We received in the first semester of 2015 **two high performance computers to enhance the Spanish chapter of DBpedia**, These machines and a high reliability data storage system will be hosted in a controlled technical room to allow 7x24 services to the scientific community and to perform high-power computations.

Partner 3: Maastricht University (UM), Netherlands

Maastricht University is the youngest Dutch university ranking in the world top 20 of universities under 50 years. The Department of Bioinformatics, BiGCaT, is a 12 year old group using bioinformatics for data integration of omics using statistical approaches and pathway analysis using pathway databases.

Relevant experience and role in project

UM develops systems biology solutions to biological questions. Central roles here are set aside for pathway databases and semantic technologies to link experimental omics data. The group has extensive experience with setting up international, multi-disciplinary data platforms, including ISATab, ToxBank, DiXA, Open PHACTS, and eNanoMapper. Experimental data is linked up into systems biology approaches using Open databases software including WikiPathways, PathVisio, BridgeDb, the Chemistry Development Kit, Bioclipse, and others. UM will focus on correctly capturing the chemical aspects of the profiles (WP2), generally contributing Semantic Web expertise, support the incorporation of two CC0 data sets (WP3), and coordination of the training events (WP5), in addition to other dissemination and communication activities (WP5).

Profile of the personnel

Dr Egon Willighagen (m) is a senior post-doc with a PhD degree in chemometrics on the computer representation and statistical analysis of chemical entities, with experience at various European institutes including the Karolinska Institutet and Cambridge University. He leads the cheminformatics toolkit the Chemistry Development Kit (CDK) and has been Invited Expert to the W3C's Health Care and Life Sciences interest group on using semantic web technologies to the life sciences. He has worked on the FP7 projects ToxBank and Open PHACTS, has been advisory board member of the compound database ChemSpider and the FP7 project OpenTox. Most relevant research to this project is the nanoQSAR platform he developed including a wiki-based knowledgebase and matching ontology, containing more than 300 ENMs and a CDK-based computational library for nanoQSAR descriptor calculation and statistical modelling building. He also collaborates with the NIH/NCBI on a semantic version of PubChem.

Prof Dr Chris Evelo (m) is the head of the Department of Bioinformatics – BiGCaT, which he founded in 2001. He was trained in Toxicology and Bioinformatics and has over 25 years of research experience. He is renowned for collaborative, open source development of bioinformatics and integrative systems biology approaches (e.g. WikiPathways, PathVisio and in Open PHACTS) and development of community standards and practical applications of those in biology. Dr Evelo coordinates the FP7 IRSES project Microgennet, he is a steering committee member of the IMI project Open PHACTS, he is a council member and coordinator for bioinformatics and systems biology activities in the NuGO foundation for nutrigenomics and its predecessor the FP6 project NuGO and the current NuGO, he is a work package leader in the FP7 systems biology programme for toxicogenomics DiXa, PI in the Dutch Consortium for Systems Biology (NCSB) and a participant in the ESFRI project EuroDish, he is an NBIC faculty and think tank and DISC coordination team member and he is a recipient of an Agilent Thought Leader Grant.

Publications

1. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. Steinbeck et al. Journal of chemical information and computer sciences 2003 43 (2), 493-500.
2. The Blue Obelisk - Interoperability in chemical informatics, Guha et al. In Journal of Chemical Information and Modeling 46 (3) 991-8
3. WikiPathways: pathway editing for the people. Pico AR et al. PLoS Biol. 2008 6(7): e184.
4. Presenting and exploring biological pathways with PathVisio. Van Iersel MP et al. BMC bioinformatics 2008, 9(1), 399.
5. Open PHACTS: Semantic interoperability for drug discovery. Williams AJ et al. Drug Discovery Today 2012, 71: 21/22. 1188-98 Nov.
6. Scientific lenses to support multiple views over linked chemistry data. Batchelor C et al. The Semantic Web – ISWC 2014. Vol. 8796 of Lecture Notes in Computer Science. Springer International Publishing, pp. 98-113.

Relevant previous projects or activities

1. eNanoMapper - EC project that services a cluster of more than 30 FP7 and H2020 projects in the area of safety of nanomaterials. It develops an ontology and database platform to support the ongoing academic and industrial research and development.
2. Open PHACTS - IMI-funded project to develop a Semantic Web platform to support research in the pharmaceutical industry.

Partner 4: Wikimedia Deutschland (WMDE), Germany

Wikimedia Deutschland – Gesellschaft zur Förderung Freien Wissens – is a charitable non-profit organisation under German Law. Founded in 2004, it is the first and largest chapter of the Wikimedia Movement, as strong partner of the US-based Wikimedia Foundation and a leader in the free knowledge movement with a membership of 22000+, 70 employees and an annual budget around €5M. WMDE supports Wikimedia projects such as the German-speaking Wikipedia, Wikimedia Commons, Wikisource and others.

Relevant experience and role in the project

In 2012, WMDE began developing Wikidata, a free, collaborative, multilingual, secondary database, collecting structured data to provide support for Wikipedia, Wikimedia Commons, and other Wikimedia projects. Wikidata has quickly evolved into a structured data repository with many uses beyond the Wikimedia projects. Since 2012, Wikidata has quickly gathered a community of more than 1.5 million editors who have so far contributed about 190 million edits to Wikidata, resulting in the creation of about 13 million data items. In addition, the potential of beyond-Wikimedia uses has attracted the attention of many individuals, groups and stakeholders, including the research and science communities. The “Gene Wiki” and the Wikidata community recently announced that every human gene

(according to the United States National Center for Biotechnology Information) now has a representative entity on Wikidata. Next steps to support life sciences data in Wikidata are to establish bots that populate Wikidata with entities representative of two other key classes: diseases and drugs; and to expand the scope of these bots to include the addition of statements that link these entities together into a valuable network of knowledge.

In November of 2014, Wikidata received the Open Data Institute's Open Data Publisher Award - celebrating "high publishing standards and use of challenging data". In December of 2014, Google announced that all the data in Freebase will be migrated to Wikidata and that Freebase will be wound down: "We believe strongly in a robust community-driven effort to collect and curate structured knowledge about the world, but we now think we can serve that goal best by supporting Wikidata -- they're growing fast, have an active community, and are better-suited to lead an open collaborative knowledge base." With this, Wikidata has become the premier structured data base project, uniquely positioned to serve as a repository for research data as well.

Wikimedia Deutschland will be the leader for WP4, Enabling the use of Wikidata in research contexts. WMDE will support the project in developing a SPARQL endpoint to enable easy access to the data in Wikidata for researchers and make Wikidata follow established standard procedures of the semantic web.

Profile of the personnel

Lydia Pintscher (f) is the product manager for Wikidata at Wikimedia Germany. She has extensive experience in Open Source and Open Content projects. She studied computer science at the Karlsruhe Institute of Technology with a focus on open collaboration, language and medicine. After that she joined Wikimedia Germany to manage the community side of the development of Wikidata and later moved on to doing product management for Wikidata. She is a regular contributor to many Open Source projects like KDE, where she is the current president of KDE e.V., the non-profit supporting the KDE community.

Daniel Kinzler (m) works as a software developer and system architect of the Wikidata project at Wikimedia Germany. Daniel wrote his diploma thesis in informatics about extracting a multilingual thesaurus from Wikipedia, and has been involved in several projects related to semantic integration and mining of semi-structured data since. During his employment at Wikimedia Deutschland since 2008, Daniel has had several roles ranging from software development to serving in the MediaWiki architecture committee.

Tobias Gritschacher (m) is part of the Wikidata project since it started in April 2012. His responsibilities ranged from software development to browser testing and test automation. Currently he is the SCRUM master for the Wikidata team and other software development teams at Wikimedia Deutschland. He studied Software Development and Business Management at Graz University of Technology. He has previously worked at crowd-sourcing and crowd-funding platforms Neurovation.at and 1000x1000.at as well as within

the Open Source project Catrobat.org which provides free open source educational apps for children and teenagers.

Publications

1. Erxleben F, Günther M, Krötzsch M, Mendez J, Vrandečić D 2014. Introducing Wikidata to the Linked Data Web. In Proceedings of the 13th International Semantic Web Conference. Springer.
2. Krötzsch M. 2014. How to use Wikidata: Things to make and do with 30 Million Statements - Wikimania 2014
3. Wikibase software

Relevant previous projects or activities

1. EU FP7 RENDER - Reflecting Knowledge Diversity
2. First phase of Wikidata development (funded by Google, AI² and the Gordon and Betty Moore Foundation)
3. Second Phase of Wikidata development (funded by Yandex and Wikimedia Deutschland)
4. Wikidata meets Archeology
5. Upload of images of the Bundesarchiv including mapping of data

Significant infrastructure

Wikimedia Deutschland's offices are located in Berlin, Germany and house 70 employees, three program departments, an executive office, finance department, communications, fundraising and evaluation departments. The organization features all necessary technical and physical infrastructure, as well as programmatic and financial management capacity to effectively participate in the Wiki4R consortium. The technical infrastructure of Wikidata is provided by the Wikimedia Foundation. This includes hosting of the website itself, bug tracking, source code repositories and more. Office infrastructure for the Wikidata development team is provided by Wikimedia Germany. This includes computers, administration and facilities.

Partner 5: Universitat Oberta de Catalunya (UOC), Spain

The Universitat Oberta de Catalunya (UOC, Open University of Catalonia) is a state-of-the art technological university with a highly innovative learning model, providing a benchmark for quality in both teaching and R&D. It was created in 1994 as one of the world's very first completely online higher education establishments and currently has more than 50,000 students. The UOC's core goal is to be the university of the knowledge society, promoting innovative education, personalised learning, technological leadership, R&D work on the information society and e-learning and the dissemination of knowledge. The UOC promotes R&D activities via 45 groups linked to a department or to one of the university's research centres: the eLearn Center, devoted to e-learning studies, and the Internet Interdisciplinary Institute (IN3), specialising in the study of the networked society and the

knowledge economy, network technologies and specific software development. In total, more than 400 people work in R&D at the UOC.

Relevant experience and role in project

Over the last five years, the UOC has participated in more than 260 R&D projects, either national or European. What is more, the UOC works to promote knowledge transfer and has, over the last four years, signed more than 1000 agreements to this end. The UOC forms part of more than 30 international networks, including the European University Association (EUA), the International Council for Open and Distance Education (ICDE) and the IMS Global Learning Consortium. The research group involved in this proposal, Open Science and Innovation, belongs to the above mention IN3, and has recently carried out a research project on the use of Wikipedia at universities. At present the group develops another research project on Science & Wikipedia which aims to analyse the scientific contents of Wikipedia and to promote the active contribution to Wikipedia by scientists and researchers. UOC's role in the project will be to develop and organize the training activities aimed at fostering the use of Wikidata for research among different potential users (WP5). This will mainly encompass the design and development of tutorials and course materials able to be used both in e-learning environments and in traditional educational settings, and the organization of training events for different audiences.

Profile of the personnel

Eduard Aibar (m) is an associate professor at the Department of Arts and Humanities and a researcher in the Internet Interdisciplinary Institute, both at the Universitat Oberta de Catalunya (Barcelona, Spain). He teaches Science and Technology Studies and has published several works on the interaction between technological innovations and social change in different arenas. He leads the research group on Open Science and Innovation and has recently finished a research project on the use of Wikipedia by faculty members in higher education and is leading a new project, funded by the Spanish Ministry of Education and Science that aims to analyse the scientific content of Wikipedia.

Josep Lladós (m) is an associate professor at the Department of Economics and Business Studies and a researcher in the Internet Interdisciplinary Institute (IN3), both at the Universitat Oberta de Catalunya (Barcelona, Spain). He teaches economic geography and international economics and has published several works on innovation processes. He is a member of the IN3's research group Observatory of the New Economy, focused on the study of the new forms of digital businesses. He has recently participated in a research project on the use of Wikipedia by faculty members in higher education and is taking part in a new project, funded by the Spanish Ministry of Education and Science, which aims to analyse the scientific content of Wikipedia. Currently he is the director of the Internet Interdisciplinary Institute (IN3).

Antoni Meseguer-Artola (m) is an associate professor at the Department of Economics and Business Studies and a researcher in the Internet Interdisciplinary Institute (IN3), both at the Universitat Oberta de Catalunya (Barcelona, Spain). He teaches Statistics,

Econometrics and Mathematics and has published several works on price competition on the Internet, driving factors to e-commerce diffusion, consumer behaviour on virtual learning environments, game theory, and e-learning. He is a member of the IN3's research group Observatory of the New Economy, focused on the study of the new forms of digital business, and the relationship with the online consumer. He has recently participated in a research project on the use of Wikipedia by faculty members in higher education and is taking part in a new project, funded by the Spanish Ministry of Education and Science, which aims to analyse the scientific content of Wikipedia.

Julià Minguillón (m) received his Ph.D. degree from the Universitat Autònoma de Barcelona (UAB) in September 2002. In January 2001 he joined the Universitat Oberta de Catalunya (UOC) where he is a faculty member of the Computer Science, Multimedia and Telecommunication Studies department. He has developed learning resources for object oriented programming, abstract data types engineering and compiler construction. He is also involved in the integration of e-learning standards in virtual learning environments, such as IEEE LOM, SCORM and IMS LD. His main research interests include the formal description of the learning process by means of ontologies, personalizing the learning process by means of adaptive itineraries based on reusable learning objects, and user modelling in virtual e-learning environments applying web and data mining techniques for improving user experience and usability, accessibility and mobility issues. He is also interested in open educational resources and the uses of social tools such as Wikipedia for teaching and learning.

Publications

1. Meseguer A, Aibar E, Lladós J, Minguillon J, Lerga M (forthcoming). “Factors that influence the teaching use of Wikipedia in Higher Education”. Journal of the Association for Information Science and Technology.
2. Aibar E, Lladós J, Minguillon J, Meseguer A, Lerga M (forthcoming). “Wikipedia at University: what faculty think and do about it”. The Electronic Library. Vol. 33. Issue 4.
3. Aibar E 2014. “Lessons from the Digital Divide”. In: Antonio López Peláez (ed.). The Robotics Divide. A New Frontier in the 21st Century? New York: Springer; 157-171. ISBN: 978-1-4471-5358-0
4. Aibar E 2014. “Ciència oberta, encerclament digital i producció collaborativa”. In: T. Iribarren, O. Gassol and E. Aibar (eds.). Cultura i tecnologia: els reptes de la producció cultural en l’era digital. Lleida: Punctum; 99-120. ISBN: 978-84-9419874-8.
5. Aibar E 2013. “Producción colaborativa y ciencia: un estudio empírico sobre las percepciones y prácticas del profesorado universitario respecto la Wikipedia”. A: González Alcaide, G., Gómez Ferri, J. i Agulló, V. (eds.). La colaboración científica: una aproximación multidisciplinar. València: Nau Llibres; 381-392. ISBN 978-84-7642-930-3

Relevant previous projects or activities

1. The use of Internet open content for university education: an empirical study on the perceptions, attitudes and practices of university faculty on Wikipedia. Recercaixa: ACUP/Fundació La Caixa; from 1-1-2012 till 31-12-2013. Ref: 2011ACUP00051
2. Information, culture and knowledge: New citizens' practices, new public policies. A case comparison of Spain and United States (Sinde Law versus SOPA Law). Ministerio de Economía y Competitividad (Spain). Ref: CSO2012-37851. From 1-2-13 till 31-1-2015.
3. Scientific authority in the public sphere in twentieth-century Spain. Ministerio de Economía y Competitividad (Spain). Ref. HAR2012-36204-C02-02. From 1-2-2013 till 31-2-2015.
4. iCity: Linked Open Apps Ecosystem to open up innovation in smart cities. Programme: Competitiveness and Innovation Framework Programme (CIP) - The Information Communication Technologies Policy Support Programme (ICT-PSP). Ref. 297363. From 2013 to 2015.

Partner 6: Europeana Foundation (EF), Netherlands

The Stichting Europeana (Europeana Foundation) is a foundation under Dutch law that owns and operates the Europeana Digital Service Infrastructure (DSI). The Europeana Foundation aims to transform the world with culture. Europeana is the initiator of a network, representing more than 2500 cultural heritage organisations and a thousand individuals from these and other walks of life, passionate about bringing Europe's vast wealth of cultural heritage to the world. Doing so will unlock untold economic and societal benefits, transforming lives in the process. Culture unites Europe, and making it more accessible promotes understanding and new economies.

The Europeana Foundation's responsibilities include providing a legal framework for the governance of Europeana DSI, employing staff, bidding for funding and making sure the Europeana DSI is sustainable.

Relevant experience and role in project

Europeana has collected and created structured information (metadata) about the objects held in Europe's combined collections, made available through a single interface. Europeana developed data standards – specifically the Europeana Data Model (EDM) – to make that information interoperable on the web, and shares that information as widely as possible by applying the Creative Commons Public Domain Mark to all of its metadata. At the same time, Europeana has started to engage users very personally in their shared history through collection days across Europe, Europeana 1914-1918 (the largest repository of personal stories about the First World War) and Europeana 1989 (one of the largest collections on the events of 1989 in Central and Eastern Europe).

Within this project, the Europeana Foundation will both make data available and explore its usage via Wikidata. Two Linked Open Data sets will be made available, Firstly, 30 million metadata records related to digitised cultural heritage across Europe, and secondly a set of 90 million bibliographic records collected by The European Library, a subunit within the Europeana Foundation concentrating on the library sector. Once this data is made available on Wikidata, The Europeana Foundation will work with project partners to explore different angles by which the data can be used both by the digital humanities research community and the cultural heritage domain.

Profile of the personnel

Jill Cousins (f) is Executive Director of the Europeana Foundation and Director of The European Library. She created both operational services, The European Library and Europeana. She has a strong web publishing background, having worked for VNU as their European Business Development Director and then transferred the lessons learnt from commercial business-to-business publishing to scholarly publishing working for Blackwell Publishing and several other academic publishers in the UK. Prior to a publishing career, she worked in the online environment for many years, first as a researcher with her own company specialising in providing business information to large corporate companies. After selling this company Jill worked as the Marketing Director for Online information. She has been involved in several international publishing industry bodies such as CrossRef and COUNTER.

Antoine Isaac (m) is R&D Manager at Europeana Foundation. He holds a PhD in Computer Science from University Paris-Sorbonne, where he started to work on applying Semantic Web and Linked Data techniques for cultural heritage (then at the French National Audiovisual Institute, INA). He has contributed to various national and European research projects, and has been involved in a number of W3C groups, notably for SKOS and Library Linked Data. He is co-author of the French book “Le Web sémantique en bibliothèque”. He is also a guest researcher at the Web & Media group in the Free University Amsterdam.

Alastair Dunning (m) is currently Scientific Coordinator of the Europeana Cloud project, and has several years’ experience running and being involved in large-scale projects at a European level, such as Europeana Newspapers and Arrow Plus. He has particular interest in how knowledge from specific academic disciplines is shared with other disciplines. At his previous job for the UK funding Agency JISC, he initiated a portfolio of projects related to harmonising digitised metadata and content from different disciplines to allow for greater cross searching and reuse.

Victor-Jan Vos (m) is Head of Programmes, Policy and Research at the Europeana Foundation. He has been working with Europeana Foundation since May 2014; before that he held different positions in collection development, online services and digital preservation at the Koninklijke Bibliotheek, the National Library of the Netherlands. He holds a master in Media Studies from the University of Amsterdam.

Publications

1. Wickett M. K, Isaac A, Doerr M, Fenlon K, Palmer C, Meghini C. Representing Cultural Collections in Digital Aggregation and Exchange Environments. D-Lib Magazine, 20(5-6), 2014.doi:10.1045/may2014-wickett
2. Stiller J, Petras VMaria G, Isaac A. Automatic Enrichments with Controlled Vocabularies in Europeana: Challenges and Consequences. Proceedings of the 5th International Conference on Cultural Heritage (EuroMed 2014). http://link.springer.com/chapter/10.1007%2F978-3-319-13695-0_23
3. Antoine I, Haslhofer B. Europeana Linked Open Data -- data.europeana.eu. Semantic Web Journal, 4(3):291-297. Wang S, Isaac A, Charles V, Koopman R, Agoropoulou Ai, van der Werf T. Hierarchical structuring of Cultural Heritage objects within large aggregations. Proceedings of the 3rd International Conference on Theory and Practice of Digital Libraries (TPDL 2013). http://arxiv.org/abs/1306.2866
4. Alastair D, Gregory I and Hardie A. Freeing up digital content with text mining: new research means new licences. Serials, 2009, vol. 22, n. 2, pp. 166-173. http://eprints.rclis.org/18049/

Relevant previous projects or activities

1. EU ICT-PSP - Europeana Cloud. Europeana Cloud is a three year project to explore cloud infrastructures in cultural heritage field. pro.europeana.eu/web/europeana-cloud
2. EU ICT-PSP - Europeana Creative. Europeana Creative is a European project which enables and promotes greater re-use of cultural heritage resources by creative industries http://www.pro.europeana.eu/web/europeana-creative
3. EU ICT-PSP - Europeana Versions 1, 2 and 3 – three projects that have provided funding for the core functions of European

Partner 7: Université Paris Sud (UPS), France

Université Paris Sud is one of the largest research universities in France, with a large spectrum of scientific disciplines and fields. It has a leading role in physics and mathematics; driven by its close ties with chemistry and biology, pharmaceutical research at Paris-Sud University focuses primarily on therapeutic innovation, active principles, therapeutic target identification, and drug vectorization. Medical research, be it fundamental, clinical or translational, is behind Paris-Sud University's leading role in key areas such as oncology, immunology and biotherapy, neuroscience and reproductive endocrinology, and public health; research in social sciences include law, economics and management.

Université Paris Sud is a co-founder of the Center for Data Science (CDS) at Paris Saclay. The goal of this initiative is to establish an institutionalized agora in which these scientists can find each other, exchange ideas, initiate and nurture interdisciplinary projects, and share their experience on past data science projects. To foster synergy between data

analysts and data producers we propose to provide initial resources for helping collaborations to get off the ground, to mitigate the non-negligible risk taken by researchers venturing into interdisciplinary data science projects, and to encourage the use of unconventional forms of information transmission and dissemination essential in this communication-intensive research area. The CDS fits perfectly in the recent surge of similar initiatives, both at the international and at the national level, and it has the potential to make the University one of the international forerunners of data science.

Relevant experience and role in project

The Data group at CDS develops a Data as a Service platform which goals and technologies are closely related to the project. The future platform will interconnect all the data of laboratories amongst themselves but also with the scientific information on the Web. Wikidata has already become a major focus point to openly share scientific information on the Web and so, it will play a major role in the future platform of CDS.

CDS includes the Machine Learning and Optimization team (AO), a joint team of INRIA/CNRS/University Paris Sud. The team has been participating and taking a leading role in several EU-projects (PASCAL, SYMBRION, CITINES, MASH, EGEE/EGI) as well as in national industry-oriented projects such as TIMCO (Technologies for In-Memory Computing).

Its role in the project will be to experiment the manners to reuse and to complete the information of Wikidata in the tools of laboratories (WP4). The aim is to start to build in the university a unified research area open to the world. Moreover, it will establish links with the ongoing effort of the EGI.eu foundation towards open science as described in the Open Science Commons whitepaper.

Profile of the personnel

Dr. Cécile Germain (f) is a full professor of Computer Science. She is a member of the Machine Learning and Optimization team (AO), a joint team of INRIA/CNRS/University Paris Sud. Her research interests are in Machine Learning and its applications to e-science and autonomic computing. As the policy officer of University Paris Sud for Scientific Computing and chair of the Data group in CDS, she has extensive experience with the scientific data systems at all scales, including international scientific collaborations in High Energy Physics. Current relevant projects are the Center for Data Science, EGI.eu and TIMCO. She has initiated and leads the Grid Observatory initiative (http://www.grid-observatory.org), a digital curation facility for the digital assets of the EGI flagship European grid for computer science and engineering. She is also involved in the design of interdisciplinary scientific challenges, including the recent HiggsML challenge.

Karima Rafes (f) is a researcher at the Center for Data Science of Paris-Saclay and she started a thesis related to Linked Data. Her expertise is research about the Semantic Web since 2007. She developed the LinkedWiki tool for MediaWiki; and in the CDS project, she researched pragmatic solutions to help laboratories integrate the latest Cloud and Linked Data technologies in their workflows.

Publications

1. Rafes K, Nauroy J, Germain C 2014. TFT, Tests For TripleStores. Semantic Web Challenge 2014.
2. Rafes K (ongoing). SPARQL Protocol and RDF Query Language (course on Wikiversity, in French).
3. Feng D, Germain C, Glatard T 2013. Efficient distributed monitoring with active collaborative prediction Future Generation Computer Systems, 29(8).
4. Zhang X, Furtlehner C, Germain C, Sebag M 2014. Data Stream Clustering with Affinity Propagation. IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 26, no. 7, pp. 1644-1656.
5. Germain-Renaud C, Cady A, Gauron P, Jouvin M, Loomis C, Martyniak J, Nauroy J, Sebag M 2011. The Grid Observatory. 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid) pp 114-123.

Relevant previous projects or activities

1. EU FP7 NoEs PASCAL and PASCAL II, Pattern Analysis, Statistical Modelling and Computational Learning
2. EU FP7 CP-CSA EGI-inspire, Integrated Sustainable Pan-European Infrastructure for Researchers in Europe
3. Kaggle challenge, Higgs Boson Machine Learning Challenge

Significant infrastructure

The Virtual Data computing centre located at the Orsay campus of University Paris Sud is a joint facility with other regional institutions. This top-level infrastructure (PUE 1.3) is currently designed for 400KW IT, and will be extended up to 1.5MW IT in the next seven years. It hosts ~ 4000 cores and 500TB of disk storage. It includes a tier2-node of the EGI grid and a 1000 cores cloud using the IaaS technology developed in the FP7 StratusLab project. Extensions are planned in 2015-2016 to include resources dedicated to database operations (project OpenData@UPSud), with 100TB of storage and in-memory computing facilities.

Third parties involved in the project (including use of third party resources)

No third parties are involved.

Associate Partners

List of associates involved as key stakeholders in the project. Their letters of support are attached in Section 6. Further stakeholders will be involved during the project.

Ethics and Security

Ethics

There are no specific ethical issues associated with the Wiki4R project.

Security

Please indicate if your project will involve:

- Activities or results raising security issues: **NO**
- 'EU-classified information' as background or results: **NO**

Acknowledgements

This proposal received a lot of support from people not listed amongst its authors. Carla Pinho and Nicola Zeuner helped with the administrative and budgetary aspects, while many others contributed anonymously or wish to remain so. We would like to extend our sincere thanks to all of them.

Funding program

EINFRA-9-2015: e-Infrastructures for virtual research environments