An adaptable domain-specific dissemination infrastructure for enhancing the visibility of complementary and thematically related research information

Engin Sagbas;[1] *York Sure*[1, 2]

[1] GESIS – Leibniz Institute for the Social Sciences, Information Processes in the Social Sciences (IPS), Bonn, Germany
[2] University of Koblenz-Landau, Institute for Computer Science, Information Systems and Semantic Web (ISWeb), Koblenz, Germany
{engin.sagbas, york.sure}@gesis.org

Abstract

We introduce an adaptable domain-specific infrastructure for dissemination of heterogeneous outcomes (e.g. publications) from thematic complementary and related projects. Our aim is to enhance the visibility of thematically related research information and to face obstacles from both sides: needs from information users and information providers. Users are confronted with finding sources for relevant information, handling with heterogeneous information display, varying information granularity on different sources, extracting and compiling the information found whereas information providers have costs for implementing and maintaining such an infrastructure from scratch, limit or omit coupling with different related sources and offer information partly in an interconnected manner. The contributions of this paper include a model closely related to the CERIF standard and a technical infrastructure ready to reuse to set up a research information system for a new research topic. We created a reference portal on the topic "Governance in the EU".

Keywords: dissemination infrastructure; information retrieval; research information; CERIF

1. Introduction

Information visibility of complementary and related information on the web is an important claim from a user's perspective. For example, collaborative research in large projects and complementary research by other related projects across national and international research institutes have the problem not to be adequately visible for those who are not familiar with the related projects. Transparency is hampered, e.g. about the produced outcomes in a research field, established research structures and connections to other related projects. Furthermore, there are different target groups with different information needs like researchers seeking new, relevant papers; project coordinators and managers looking for project specific documents; and the general public interested in new developments in specific research topics. Users with these different needs usually start finding the relevant information by using different search engines or available specific project information systems. It is a very tedious and time consuming task for a user to find and use several relevant information systems and websites. The success of finding the requested information across several sources is uncertain. In addition, the results found are heterogeneous, i.e. they mostly have a different kind of information display and granularity. Besides, if the information is not directly interconnected to related sources, the access to relevant complementary resources is hampered. For the information provider there is a challenge to build such a project dissemination infrastructure usually from scratch that gathers these information needs from the user.

2. Challenges

We identify challenges from two sides: On the one side, the information consumer needs, and, on the other side, the information provider needs. The information consumer side usually consists of individuals participating in the projects, users of the projects' results like external researchers, policy makers, and the interested public. They are confronted with the following obstacles:

- In project information systems like CORDIS (European Commission's Research Information System) [1], information is currently available at the level of the individual research projects. Persons interested in individual project outcomes like conferences and publications are required to visit the websites of all projects dealing with the topic of interest. The visibility of the projects and of their collective contribution to the realization of the Framework

Programme priorities and European Research Area [2] is therefore rather limited.

- Users who will not know in advance which type of information or service is to be expected from each project website, are forced to find and visit all project websites including those not relevant to them.
- Due to the lack of interconnections to complementary and related information across project boundaries, users will usually have to visit multiple websites for further information needs.
- By visiting each website, users have to learn the sites' structure, how to find and access relevant information on each project website, and finally compile themselves the heterogeneous materials with varying qualities found on different sites [3]. Mostly, it is a time consuming and inconvenient task for users.
- Biased by the above problems, users miss the big picture for relevant and related information.

In contrast, the information provider needs are characterized with the following common situations:

- Spending time and financial resources for implementing the project dissemination activities for each project resulting in several websites with similar infrastructures which are usually project-specific and isolated. Therefore, they are not coupled with the complementary information from other information providers.
- Across all projects, different dissemination and sustainability strategies beyond the projects' duration will make it difficult to ensure the availability of project results in the long run yielding information websites that are not maintained or no more visible [4].
- The lack of a topic-oriented research infrastructure for dissemination of complementary project outcomes can lead to an unnecessary duplication of work on the provider side, and an increased effort for finding and accessing relevant information on the user side.

3. Approach

Our contributions are making thematically connected research activities visible at a single place together with their results, giving users integrated access to currently distributed resources at a common level of quality of service; to provide an adaptable technical infrastructure for information providers facilitating dissemination of heterogeneous outcomes from thematic complementary projects targeted to different audiences; to integrate and

compile heterogeneous data from different sources providing quality data for the purpose of analyzing, visualizing and reusing by other services; to provide a collaborative infrastructure connecting interested and active researchers; to reuse the complete information infrastructure for a new domain reducing costs for acquisition and implementing; and to facilitate sustainability of project outcomes after the project's end.

The main pillar of work carried out focused on the development of a technical dissemination infrastructure which covers all entities relevant in the context of research information (i.e. actors, activities and results) at a very detailed level and at the same time interconnected them both within the context of an individual project and across project boundaries.

3.1 The Conceptual Model

In the first phase of the EU project IConnectEU [5], the different outcomes produced by eight complementary projects and the target audience of these outcomes were analyzed and a core model was defined for documenting these outcomes together with information about participating institutes and persons at a very detailed level. The core model is closely related to the Common European Research Information Format (CERIF), which was funded by the European Commission and is maintained by euroCRIS [6], a professional organization dedicated to improvement of research information availability since the release of CERIF2000 [7]. Compatibility to CERIF, in specific to its exchange format CERIF-XML [8], supports reusing research information across institutional and geographic boundaries.

The CERIF model is built around three core entities of research information and three result entities. Core entities are project, person and organizational unit. The result entities consist of publication, patent, and product. These entities are connected with typed links which represent the semantic relationships between these entities expressing, e.g. the members of a project, the affiliation of a person, project outcomes, authors of publications, and persons with specific project roles like coordinator, to name a few examples.

These entities are reflected in the information architecture, where semantic annotations, i.e. attributes, were used to describe these entities. They have been partly expanded in regard to the attribute set defined in CERIF. This not only includes additional information on e.g. project work packages, data collections or scientific methods, but also includes geographic location and coverage of all entities, target groups of activities and results.

We specified the basic requirements in a core model that includes all relevant entities, with their describing attributes and the relationships

between them. We modeled the research information context comprising of the following core entities: project, institute (institutional participants involved in a research project), person (doing research and affiliated with an institute), and research results (including project research outcomes like publications, events, research data or other produced results).

Figure 1 depicts the conceptual model. Each of these entities (displayed in oval form) has its own set of mandatory and optional attributes, which adequately describe the single entity. Attributes are grouped in formal attributes, specific attributes, attributes for content describing and indexing, and finally, semantic relations for interconnecting entities.

For example, attributes for describing projects are divided in:

- formal attributes: title, acronym, begin, end, funding, URL etc.,
- specific attributes: funding agency, thematic priority, instrument etc.,
- content describing attributes: summary, research area, objectives, work packages etc.,
- content indexing attributes: geographic coverage, thesaurus keywords, free keywords etc., and
- semantic relations: linking semantically the project entities to institutes, persons, and results and interconnecting all entities in the core model to express relationships between two entities, e.g. “Mike” (person) is involved in “Apollo” (project). Other relation type attributes can be defined and used like “coordinator of”, “author of”, or “cooperate with”.

The result entity is subdivided in different research outcome entities produced by all projects. Especially publications are the most prominent example for representing research outcomes. Events like conferences or workshops, produced results like research data, and other generated project resources are also covered project outcomes. Each result entity has its own describing attribute set and is interconnected within the research context.

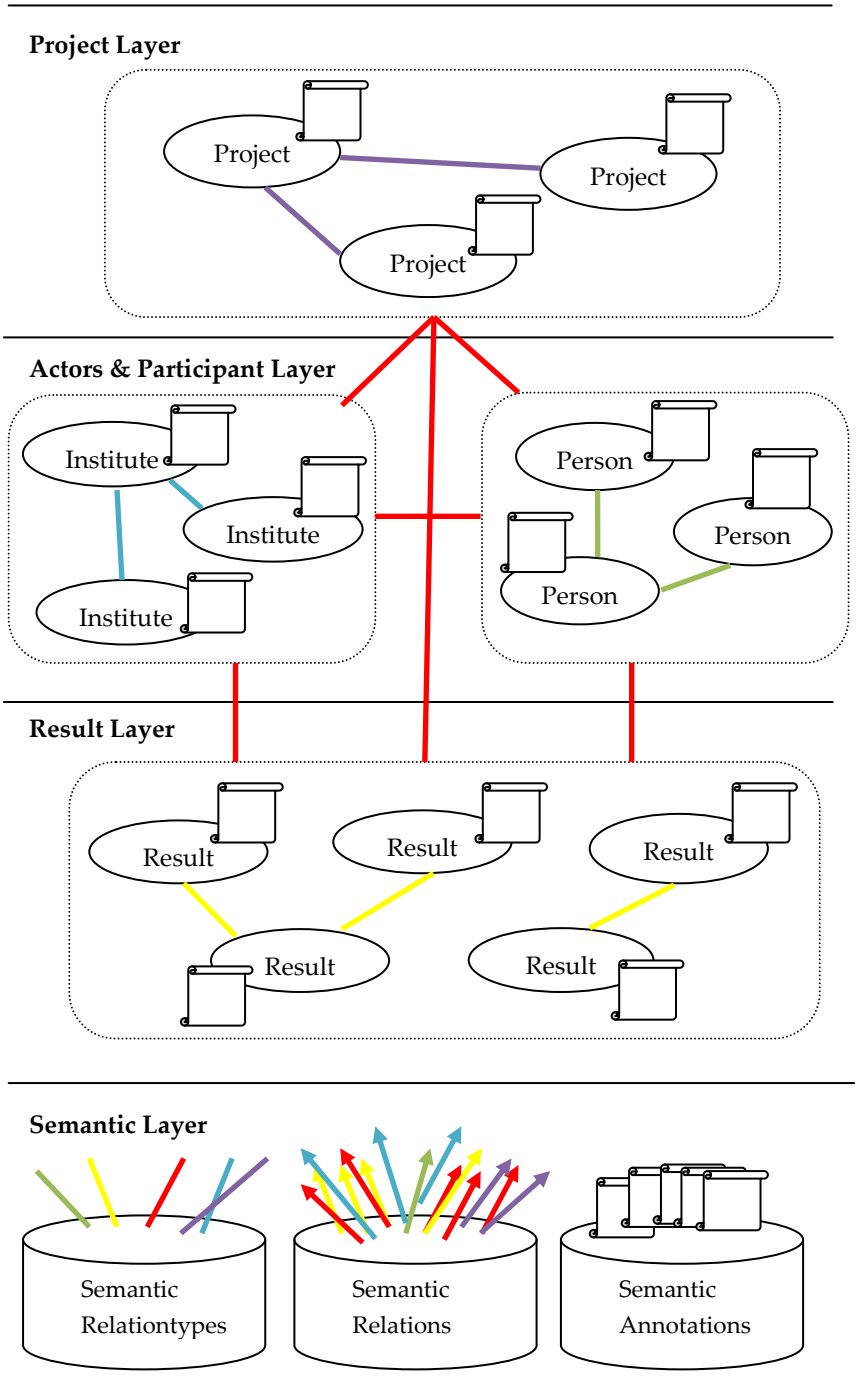

Figure 1: Conceptual model in the context of research information

3.2 The Architecture Model

The overall architecture model is shown in Figure 2. The model considers the different information needs and implements the conceptual model described in the last section. The middleware consists of two main parts. A content management system manages the editorial static contents. The dynamical contents representing the conceptual model in Figure 1 are realized by a cataloguing system. This combination allows exploiting the synergy effects of both specialized systems so that information providers can both build complex portal structures combined with functionality of a cataloguing system providing dynamic contents.

We use for the technical infrastructure of the IConnectEU reference portal as cataloguing system DBClear [9], which is developed in a project funded by the German Research Foundation (DFG). Due to its flexibility, DBClear has successfully been adapted to several use cases where a web-based cataloguing system was needed to collect and map information, e.g. in the FP5 project "MORESS - Mapping of Research in European Social Sciences and Humanities", in around 10 Digital Libraries in Germany, and recently in "SSOAR – Social Science Open Access Repository"[10], which was also funded by the DFG. As content management system we adopted Typo3 [11]. Both software packages are open source.

Our infrastructure is flexible in regard to adapting it to a new domain. Since the conceptual model developed is generic for research information, it is not restricted to a particular research discipline. All kinds of research topics in different research domains can in principle be covered by the model, e.g. research topics in social sciences, life sciences, or natural sciences. Reusing the conceptual model can significantly reduce the effort to set up a specific topic-oriented research information system.

Customizing and extensions might be useful to adapt the model to new emerging needs, e.g. for particular requirements of a new domain. DBClear has two main strengths. The flexibility both in defining and editing semantic annotations and flexibility in adapting the information display view for the user interface, both feasible during and after system implementation. For example, we could define a new relation type called "cooperate-with" and then annotate persons who cooperate with each other (semantic relations). This would represent a cooperation network of persons. There are no limitations, i.e. we can add or adapt all semantic annotations and semantic relation types for the given entities in Figure 1.

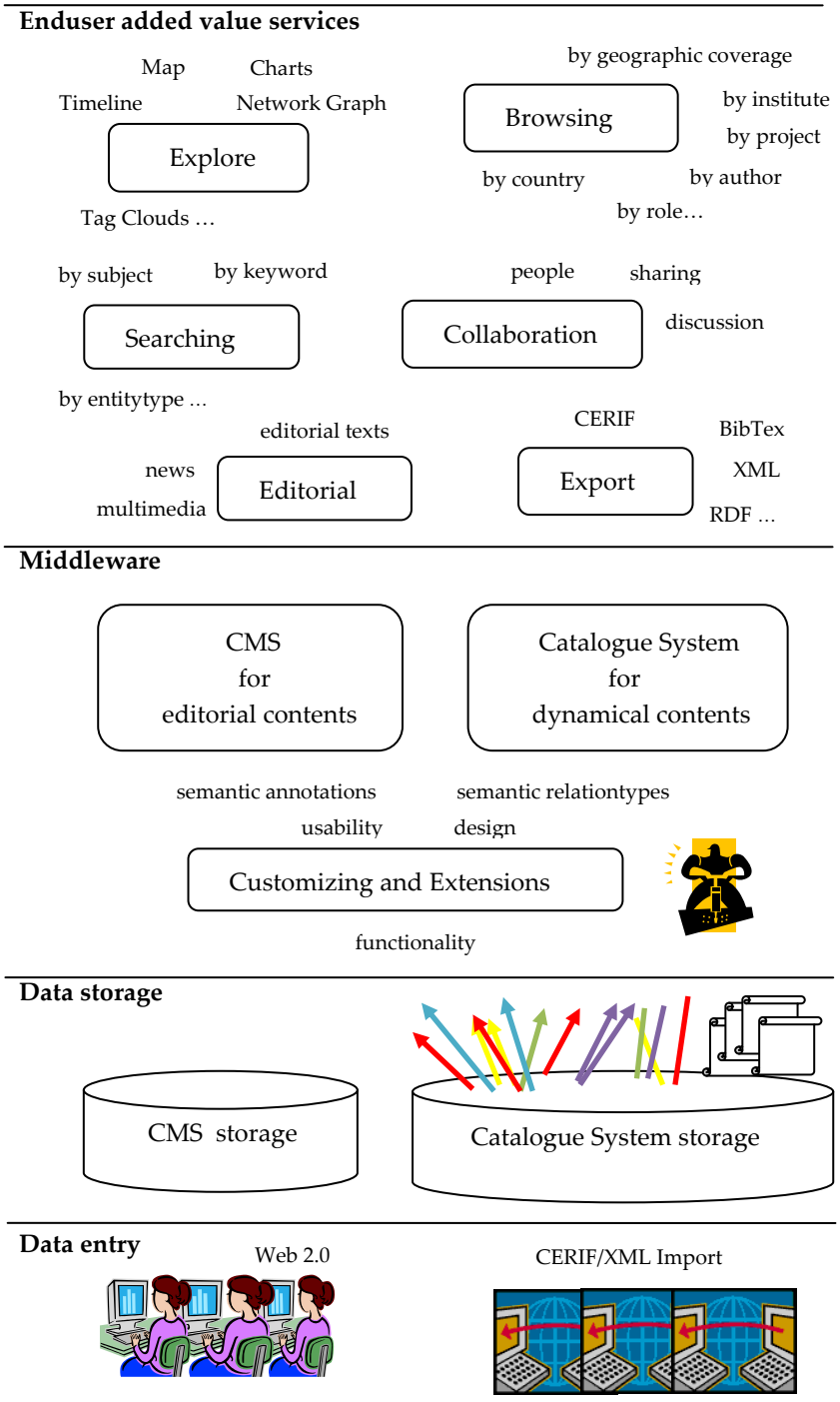

Figure 2: Overall architecture model

3.3 Enduser added value services

Building a useful research information system from a user's perspective requires attention for different user needs. Especially, added value services could be a crucial incentive using it. Figure 2 lists a few valuable services. For example, browsing frequently requested research information in predefined categorized views facilitates quick access to the requested information, e.g. persons can be browsed by project members, by institute members, by project roles, by authors or by another person view adopting a relation type to a person. Further services like new visualization types (e.g. map, charts) or useful export formats for publications like BibTex can be realized.

3.4 Data entry and Data gathering

Each information system with its varying functionality and services rely on the availability of quality and up-to-date data. Thus, an issue is where to get these data. There are approaches for automatically harvesting and collecting relevant data, e.g. by web mining, but we focus here on the idea of Web 2.0. This has the advantage that a human collaboration led to better results of gathering quality data by exploiting the social intelligence [12]. We focus on researchers who would provide their own data and integrate it with the others. Another associated issue is, how we can get the user's data without forcing them to use multiple places for data entering. In this case, we can use standards for reusing research data. The CERIF standard adopting by research information systems (CRIS) makes research information across different systems available using the CERIF exchange format. Export formats from proprietary systems are imported by using XML and CSV formats. The flexibility of DBClear to define new mapping templates for different data export formats by using XSLT makes large amounts of data reusable from other quality controlled systems.

4. Related Work

The IConnectEU model is closely related to the CERIF model which is regularly updated. The current release is CERIF 2008 [13]. IConnectEU is consistently structured according to the CERIF core entities: project, institute, person, and results. There is a mapping defined for the intersection between the IConnectEU and the CERIF entities. Due to the specific use case of IConnectEU, not all entities from CERIF are relevant or used. Otherwise there are some entities and attributes defined in IConnectEU but not present in

CERIF. If required, the conceptual model can be evolved or adapted to the specific discipline or emerging new requirements.

Relying on the CERIF standard ensures that the data can be reused by other third-party research information systems. For example, the IST World project [14] adopts the CERIF standard so that the IConnectEU data can be further reused, e.g. to apply advanced technologies for visualization like a competence diagram. The IST World project complements IConnectEU since its focus is on analyzing research competencies across European countries [15]. IConnectEU's primary focus is to provide a complete model for covering typically research information combined with a ready for reuse software infrastructure applicable for any research topic.

There are other portals like sowiport.de [16] (one of the largest information portals for the social sciences in Germany) or vascoda.de [17] (an interdisciplinary portal for scientific information in Germany) which provide a broad range of information from multiple integrated databases. Science gateways like WorldWideScience.org [18] enable federated searching of national and international scientific databases and portals. In contrast to these portals, IConnectEU has a narrow thematic focus on a research topic within a discipline connecting only thematically related projects, e.g. projects with research on the topic "EU Governance". In this sense, it is a lightweight thematically focused information system not intended to be a literature database with millions of entries from different areas. Especially, IConnectEU covers the project context with research information (e.g. persons, institutes, results) in a semantically interconnected manner which is not or partly provided in this form by the mentioned portals. To sum up, IConnectEU can contribute topic oriented research information and data to larger (multi-) disciplinary portals and profit from accessibility from those portals gaining new users who have a special research focus.

Search engines like Google might be useful but their retrieval effectiveness [19] is limited in the context of finding complementary research information. Due to the fact that IConnectEU is thematically focused and the data is handpicked and quality controlled by the partners, all search results are thematically relevant and related, which result in a high retrieval quality.

5. Conclusion and future work

We addressed common needs of both information users and information providers. From the user's perspective there are obstacles to find and retrieve relevant and related research information on the web. These include finding

the relevant websites and sources, confronting with the heterogeneity on different sources like different information display and granularity, and extracting as well as compiling the collected information from the web. In contrast, obstacles for information providers are costs for implementing and maintaining a complex dissemination infrastructure, coupling with other related sources, and providing interconnected information to thematically related information. Another issue is sustainability of information beyond the projects' duration.

We introduced the benefits of the IConnectEU infrastructure allowing topic-oriented organization of complementary research information and outcomes. The typical research information, e.g. from projects, institutes and persons to publications, conferences and other results are covered and interlinked semantically. Research networks can be represented, i.e. linking researchers and institutes across projects and countries. Besides, detailed information and extensive metadata for a large number of information entities are provided. The data is compatible with the CERIF standard for research information promoted by the European Commission. Reusing data by other services for the purpose of analyzing and visualizing adds several new dimensions to geographical analysis, e.g. mobility of researchers, development of collaboration networks, and inclusion of regions in European funded research.

IConnectEU strengths are based on an adaptable infrastructure. Using it for dissemination in other research topics will significantly reduce the provider's costs. The software developed in IConnectEU is made freely available as open source software to third parties.

Future work includes the Web 2.0 approach to get research information for the data entry process directly from the involved persons. This requires a collaborative infrastructure that eases the data gathering process since everyone would be responsible for maintaining his/her own part of contribution. Our approach for dealing with the issue will include:

- Providing an incentive by establishing new added value services like new visualization and exploration services for data, e.g. map visualizations allowing geographical analysis.
- Bridging to social networks for special target groups like researchers using XING, and making an incentive to join the platform [20].

We will extent the current dissemination platform to a collaborative infrastructure where users maintain their own data and collaborate together, e.g. discussions on a topic. Further, we prove use cases for connecting to social network users by using the open social standard [21], which is also part of future work.

Notes and References

[1] CORDIS: http://cordis.europa.eu/

[2] European Commission Research http://ec.europa.eu/research/

[3] P. Oliveira, F. Rodrigues, P. Henriques, und H. Galhardas, A taxonomy of data quality problems, *Proceedings of 2nd International Workshop on Data and Information Quality*, 2005, p. 219-233.

[4] K.H. Lee, O. Slattery, R. Lu, X. Tang, und V. McCrary, The state of the art and practice in digital preservation, *Journal of Research-National Institute of Standards and Technology*, vol. 107, 2002, p. 93-106.

[5] IConnectEU: http://www.iconnecteu.org/

[6] euroCRIS: http://www.eurocris.org/

[7] A. Asserson, K.G. Jeffery, und A. Lopatenko, CERIF: Past, Present and Future: an Overview. Gaining Insight from Research Information, *6th International Conference on Current Reseach Information Systems*, 2002, p. 29-31.

[8] B. Jörg, O. Krast, K.G. Jeffery, und G. van Grootel, CERIF 2008–1.1 XML: Data Exchange Format Specification. *euroCRIS*, March, 2010.

[9] H. Hellweg, B. Hermes, M. Stempfhuber, W. Enderle, und T. Fischer, DBClear: A Generic System for Clearinghouses, *Gaining Insight from Research Information*, 2002, p. 131.

[10] SSOAR: http://www.ssoar.info/en/

[11] Typo3: http://typo3.org/

[12] J.F. Jensen, User-generated Content – a Mega-trend in the New Media Landscape, *Interactive TV: Shared Experience, TICSP Adjunct Proceedings of EuroITV2007*, 2007, p. 29–30.

[13] B. Jörg, K.G. Jeffery, A. Asserson, und G. van Grootel, CERIF 2008–1.1 Full Data Model: Introduction and Specification. *euroCRIS*, March, 2010.

[14] IST World: http://www.ist-world.org/

[15] B. Jörg, J. Ferle, H. Uszkoreit, und M. Jermol, Analyzing European Research Competencies in IST: Results from a European SSA Project, Bošnjak&Stempfhuber, 2008.

[16] Sowiport: http://sowiport.de/

[17] Vascoda: http://vascoda.de/

[18] WorldWideScience: http://worldwidescience.org/

[19] D. Lewandowski, The retrieval effectiveness of web search engines: considering results descriptions, *Journal of Documentation*, vol. 64, 2008, s. 915-937.

[20] L. Leung, User-Generated Content on the Internet: An Examination of Gratifications, Civic Engagement, and Psychological Empowerment, *New Media & Society*, 2009.

[21] OpenSocial: http://www.opensocial.org/