Entities and Identities in Research Information Systems

Brigitte Jörg[a b], Thorsten Höllrigl[c], Miguel-Angel Sicilia[d b]
[a] German Research Center for Artificial Intelligence (DFKI GmbH), Germany
[b] euroCRIS
[c] AVEDAS AG, Germany
[d] University of Alcalá, Spain

Summary

Where science is increasingly becoming a global business and furthermore with open data and access initiatives, means to identify and thus to connect system-internal with non-internal entities are urgently needed. The need for identifiers beyond systems is thus a global requirement but also relevant within organization boundaries spanning multiple systems. Various identifier initiatives and systems have started in the scientific domain and beyond. However, they have not yet achieved the required interoperability. With this paper we aim to collect additional essential *ingredients* towards designing a sustainable CERIF identifier sub model, contributing to ongoing discussions and supporting design decisions. Here, we first present common issues with identifiers in the wider academic domain, before we analyze available systems, technologies and current initiatives solving the global identity gap.

1 Introduction

Most information systems nowadays are still built upon relational database management systems (RDBMs), where tables are the physical containers for recordings of real world objects and their relationships. These have usually been designed through entity relationship models to describe the particular domain of interest, often conveying conceptual domain models (Wand and Weber 2002). “Conceptual Modeling is the activity of formally describing some aspects of the physical and social world around us for purpose of understanding and communication” (Mylopoulos 1992). Where conceptual models support well the descriptions of worlds of interest and therefore the objects within, they have so far not been much concerned with the identification of the described objects as such (Evermann and Wand 2001). Relationally built information systems identify objects through system-internal identifiers and these incorporate the referential integrity support to aggregate information contained in distributed tables converging to a real world object (e.g. a organization, person, etc.). A real world object may either be built during information integration or exchange, upon querying or through application rules, by system constraints or accumulated queries in user interfaces or pre-selected views. System-internal identifiers work well within system boundaries to identify and aggregate information about objects, but they do not scale for usage across systems. Where science is increasingly becoming a global business, research environments and thus research-related information is becoming more open and accessible; relevant information objects are replicated across multiple systems retaining sharable descriptions.

Means to identify and consequently connect system-internal with non-internal entities are therefore needed globally, but as well within organization boundaries spanning multiple systems. For example, a person and the person's first names or family names will be maintained in the human resources (HR) system, but as well, the person will be known in the project management system, the publication repository, and the financial system. In each of these systems, the *same* person object will most probably have its system-internal identifier, and in each system the name recordings may be spelling variants of records in the other (e.g. B. Jörg, Brigitte Jörg, B. Joerg). Most often, the identification of objects in information systems is dependent on institutional and legacy setups and rules (Hoellrigl et al. 2008, Bischof et al. 2007). A mechanism for identity management should therefore be in place to support the administration of various identifiers e.g. (Hoellrigl et. al. 2009), (Hoellrigl et. al. 2010).

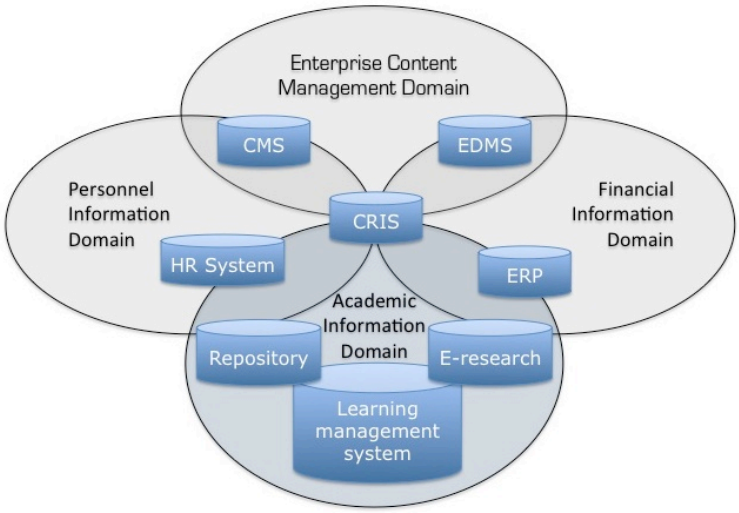

Figure 1: The Enhanced AID (Academic Information Domain) Model. (Godtsenhoven et al. 2009, p. 49)

Current Research Information Systems (CRIS(s)) have been recognized in the center of a scholarly information interoperability framework (Godtsenhoven et al. 2009); in playing a critical role with information sharing and providing the required flexibility for multiple stakeholders' needs. However, for techno-historical reasons they assumed the completeness of information within system boundaries and do not yet clearly account for various kinds of identifiers. The history of CRISs in Europe is tightly related to CERIF – the Common European Research Information Format. With CERIF[1], as an auxiliary for identifiers, the entity *cfElectronicAddress* has sometimes been employed. However, there are still ongoing discussions in the community, whether an electronic address resembles semantically the concept of an identifier, and if it therefore can or should be used as such. A CERIF model extension towards a *cfFederateIdentifier* entity has been agreed, challenging design decisions by the fact of such an entity being physically embedded in a closed information system world, but conceptually open to the world, i.e. any information system identi-

[1] The Common European Research Information Format (CERIF); a EU Recommendation to Member States: http://cordis.europa.eu/cerif/, http://www.euroCRIS.org/

fier. Recently, a CERIF model extension for linkage of CRISs through LOD entities has been proposed (Joerg et al. 2012), where further discussions are needed as to which extent the LOD CERIF extension accounts for a generic CERIF identifier sub model.

The need for multiple or global object identifiers in science systems is obvious. A scalable and thus sustainable interoperation framework or model like CERIF must account for object identification beyond and across systems. We now first present common issues with identifiers in the wider academic domain, before analyzing available systems, technologies and current initiatives aiming to solve the global identity issue. Thereby, we investigate identifier systems and initiatives in the wider scholarly domain.

2 Identifier Issues in Research Information Systems

From the introduction we have learned, that Research Information Systems in this context are not seen in the very strict sense, but take a wider scope. (Godtsenhoven et al. 2009) calls it *Enhanced Academic Information Domain* (see Figure 1). We will now demonstrate the unique identification need by presenting familiar use-cases.

2.1 Researcher Use-Case

Within an institution, researchers are typically identified via a locally unique string or number. Mostly however, this identifier is only unique within a certain system or service and often, each organizational unit, such as a computer center, a library or a university administration is generating and maintaining its own unique identifiers for each researcher with its own identifying attributes. For an aggregated view of e.g. a researcher's output, the researcher has to be uniquely identified within all contributing systems and the following challenges have to be tackled:

- First, an identifier has to be generated that is not only unique in the context of one system, but within the whole organization or even beyond organizational boundaries.
- Second, the various identities of one person have to be linked with each other.

The motivations to have organization-spanning unique person identifiers are various. The concept of an identification number that uniquely identifies a person beyond institutional boundaries is quite common in the governmental field, i.e. a national identification number. Governments use such identification numbers to uniquely identify and track citizens e.g. for the purpose of work (see also Enserink 2009, DAI). However, such identifiers – e.g. the Digital Author Identifier (DAI) in the Netherlands or the Social Security Number in Norway – are often unknown beyond national systems and typically unused by researchers. Known technologies that support usage of unique identifiers are so-called Single Sign-On (SSO) systems. They allow for personal authentication once, and subsequent use of intra- or inter-organizational services without repeated logins. Achieving SSO requires to link or to *federate* different user accounts. An open tool to support authentication based on federated SSO solutions is e.g. Shibboleth[2].

A common need of unique researcher identification is in their role as authors. Finding all publications of a certain researcher based on his given name and surname will not lead to a satisfying and

[2] http://shibboleth.internet2.edu/

complete list of all his publications, because it is not unique over all authors worldwide. It is furthermore highly possible that the surname or even a given name of a person changes. An author identification system – unique and persistent over time – is one of the most important requirements and a long-standing issue. A unique identifier could overcome error-prone record comparisons based on strings, and publications could again be linked with author identifiers. This would be extremely helpful also with data collections (e.g. CRISPool[3]) and avoid new identifier generations for record aggregations with new collections.

2.2 Publication Records Use-Case

In the research domain the unique identification of a publication is important. Once a researcher publishes a paper, he wants to make that publication available for citation, because the citation count is a common metrics about the recognition of his work. In fact, the entire research activity is built upon published results and thus previous work, and consequently, a publication should be made available for unique reference. Nowadays it is common to make publications available via the World Wide Web, with an additional requirement for long-term preservation. A system is thus required to keep the links with publications persistent; supplying both, a unique URL as well as an identifier that is cross-publisher- and system-independent (i.e. a persistent identifier). This is necessary also because a URL, which is used to link to a certain publication has to be functional even if the location of the server where the publication is stored has been moved.

Unique publication identifiers are not only supportive with citations and preservation, but furthermore with queries or exchange, and also with system migrations. Typically, different publication portals such as PubMed, a free database of the United States National Library of Medicine[4], the online portal of the German National Library (DNB)[5], Thomson Reuters Web of Science[6] or Elsevier Scopus[7] store information about publications and provide interfaces to download metadata (in some cases with restrictions). Based on this metadata, researchers can import information about their publications as well as citations. To uniquely identify one and the same publication across different online portals a common identifier is supportive, because otherwise again, record matches depend on comparisons of titles or assigned author names, which are error-prone, not unique and therefore not simply scalable.

2.3 Organisational Unit Use-Case

Assignments of unique and persistent identifiers are not only required for publications or researchers, but are also helpful with organizational unit records. Funding organizations as well as research projects (e.g. analyzing the scientific domain (Joerg et al. 2008)) or assessment exercises, would certainly profit from a globally unique organization identifier; as it would significantly improve data quality. The European Commission introduced so-called Participant Identification

[3] http://www.crispool.org/

[4] http://www.ncbi.nlm.nih.gov/pubmed/

[5] http://www.dnb.de/eng/index.htm

[6] http://wokinfo.com/products_tools/multidisciplinary/webofscience/

[7] http://www.info.sciverse.com/scopus

Codes (PICs)[8] with the Seventh Framework Programme. These are since mandatory for all organizations preparing a proposal submission. The PIC allows to uniquely identify organizations and their units, while the submitting units do not need to repeat organization-related information with each submission. The PIC number is not public.

The presented use cases are common in the research domain and the selected entities person, publication and organization, i.e. actors and output, are certainly most important in terms of research measurements, and where unique identification is surely a contribution to quality. It is important to note: global identifier systems do not merely require technological solutions but also need governance. The subsequently investigated initiatives mostly supply their own technology to implement their governing mission.

3 Globally Unique Identifiers – Initiatives and Systems

In the research world, the global identifier gap has been recognized being critical for quality improvements in information systems, and at the same time to enable large-scale information sharing or reuse. The Science journal dedicated a comprehensive article to researcher identification: *Are You Ready to Become a Number?* (Enserink 2009) and multiple international initiatives have started in this respect. Life could be much easier for involved stakeholders if research entities such as publications, organizations and people had unique identification numbers. Activities towards a global researcher identifier are proceeding at high speed. ResearcherID was the first out of the gate (Enserink 2009); a more recent approach is ORCID. Where author, contributor or researcher identifiers are of most interest now, publication identifier initiatives started more than a decade ago with the operation of CrossRef – the official DOI registration agency. At around the same time, the Virtual International Authority File (VIAF) has been initiated in the library domain, where ten years earlier, a widely distributed identifier system has started, namely Handle. Beyond the academic domain are universally unique identifiers (UUIDs), uniform resource identifiers (URIs) or the OpenID . Our list of analyzed initiatives and systems does not claim completeness. The authors tried to provide an overview of the most common and popular initiatives and systems under consideration for potential use within the research domain.

3.1 ORCID – the Open Researcher & Contributor ID

The most prominent initiative is certainly ORCID[9] – the Open Researcher & Contributor ID, which started in late 2009 "to solve the author/contributor name ambiguity problem in scholarly communications". The driving non-profit organization was created in August 2010 and membership grew rapidly. ORCID aims to launch its services in the second quarter of 2012. It licensed the Thomson Reuters ResearchID code in August 2011 and will provide a query API for:

- **Bio:** given a contributor, returning names and affiliation data
- **Works:** given a contributor, returning works they have contributed to
- **Work:** given a work, returning the contributors that are responsible for it
- **Search:** given whatever metadata, returning a ranked list of potential contributors identified by that metadata

[8] http://cordis.europa.eu/fp7/pp-pic_en.html

[9] http://about.orcid.org/

Where its phase one was concerned with setting up the service, phase two aims at disambiguation of the provided data towards a single identity employing automatic means. The ORCID initiative is financed by organizational membership fees; aimed at being free for individual researchers (Fenner 2012 slides).

3.2 ResearcherID by Thomson Reuters

Thomson Reuters's ResearcherID (http://www.researcherid.com/) launched officially in January 2008 (Enserink 2009). It aims to assign a unique identifier to each author that participates and upon which it aims to build a Researcher index. It has been endorsed for usage by (Cals and Kotz 2008), because it was the first global scheme ready and available for researcher identification. ResearcherID is currently a free service to the multi-disciplinary scholarly community, but there are concerns in the community as to how access and cost will be defined in the future. ResearcherID allows for citation metrics, searching and connecting with collaborators and *uploads* of publications from the Web of Knowledge or from EndNote Web. ResearcherID profiles store multiple name variants; the public service allows for queries by family and given names, by institution or country, and by keywords. A ResearcherID is designed as in the example: A-1026-2007.

3.3 CrossRef

CrossRef[10] started its operation in 2000 through a non-profit and independent organization formed by the world's leading scholarly publishers, namely PILA – Publishers International Linking Association. The initiative started 1999 by combining elements from two projects. Where the former was more a reference linking service using the Digital Object Identifier (DOI), the latter was more a coalition of publishers with the critical mass to launch, grow and sustain such a system, i.e. supplying a business model. The initial mission with reference linkage or DOIs has later been broadened "to enable easy identification and use of trustworthy electronic content by promoting the cooperative development and application of a sustainable infrastructure." (CrossRef 2009)

A DOI is a unique alphanumeric string and CrossRef associates with each DOI a set of basic metadata and a URL pointer to the full text on the Web. E.g. the DOI 10.1007/978-3-642-05290-3_91 through the CrossRef system (http://dx.doi.org/) is resolved to a URL allowing for access of the full text publication: http://www.springerlink.com/content/m574117014g7566l/.

3.4 Handle

Publication repositories such as DSpace[11], Eprints[12] or Fedora[13] use Handle[14] to uniquely identify publication records. The handle system includes several features to resolve handles to information

[10] http://www.crossref.org/

[11] http://www.dspace.org/

[12] http://www.eprints.org/software/

[13] http://fedora-commons.org/

[14] http://www.handle.net/

for location, access, contact, authentication or other usage of the resource. Associated information with an handle can be changed as needed, without changing the identifier itself. The Handle system has been developed by the Corporation for National Research Initiatives (CNRI), a not-for-profit organization to undertake, foster, and promote research in the public interest. The handle system infrastructure is supported by prefix registration and service fees, where the majority contributor is the International DOI Foundation. Initially, Handle aimed to develop a framework for the underlying infrastructure of digital libraries, and a little later was additionally funded by DARPA (Kahn and Wilensky 2006). Handles are being used in digital watermarking applications, GRID applications, and repositories, registries and more. Within the handle namespace, every identifier consists of two parts: the handle prefix (previously called a naming authority), and a suffix or unique „local name“ under the prefix, separated by a slash: 10.1045/january2010-reilly

3.5 VIAF – Virtual International Authority File

VIAF[15] is hosted by OCLC – the world's libraries connected, as an international service to provide access to the world's major name authority files. VIAF has been initiated by the Library of Congress, the Deutsche Nationalbibliothek, the Bibliothèque nationale de France (BNF) and OCLC. It envisions itself as a building block of the Semantic Web. Most large libraries maintain lists of names for people, corporations, conferences and geographic places, as well as lists to control works and other entities – these are called authority files, which VIAF aims to contribute and promote for re-usage. A VIAF RDF example record[16] aggregates and maps multiple naming schemes and vocabularies besides VIAF. We did not yet investigate the International Standard Name Identifier (ISNI)[17], which is primarily based on VIAF and an ISO Standard (ISO 27729).

3.6 Uniform Resource Identifiers (URIs)

The vision of the Semantic Web towards a Web beyond documents is through linkage of data by applying a basic set of principles, namely, the use of URIs as names for things. Linked data refer to URIs as global identifiers (Bizer et al. 2009). Technically, a URI is a structured string of characters used to identify a name or resource. The URI syntax consists of a URI scheme (e.g. http, ftp, file) followed by a colon character and then by a scheme-specific part. URIs can be relative or absolute, e.g. resource.txt or http://example.org/resource.txt. In fact, we are producing new URIs with every new piece of information that we store, by giving it a name.

3.7 Universally Unique Identifiers (UUIDs)

The Open Software Foundation (OSF) recommended the usage of so-called UUIDs in software construction. "Anyone can create a UUID and use it to identify something with reasonable confidence that the same identifier will never be unintentionally created by anyone to identify something else" (Wikipedia). OSF was a non-for-profit organization founded in 1988 in the US to create an open standard for implementing the UNIX operating system. In 1994, OSF announced a

[15] http://www.oclc.org/viaf/

[16] http://www.oclc.org/developer/documentation/virtual-international-authority-file-viaf/viaf-rdf-example

[17] http://www.isni.org/

new organizational model marking its end of software development. However, UUIDs are still in widespread use. A UUID is a 16-byte (128-bit) number. In its canonical form represented by 31 hexadecimal digits, displayed in five groups separated by hyphens for a total of 36 characters in the form 8-4-4-4-12, e.g.: 550e8400-e29b-41d4-a716-446655440000 (Wikipedia).

3.8 Open ID

The OpenID Foundation (OIDF) is an international non-profit organization of individuals and companies committed to enabling, promoting and protecting OpenID technologies. Formed in June 2007, the foundation serves as a public trust organization representing the open community of developers, vendors, and users, and assists the community by providing an infrastructure and help in promoting and supporting an expanded adoption of the OpenID. OpenID is a decentralized authentication protocol that makes it easy for people to sign up and access Web accounts. Among the sponsoring members are Google, Microsoft, PayPal. OpenID is not especially addressing the research domain, but operates in a general information space.

4 Analysis and Verification

The presented initiatives have similar goals – aiming at globally unique identifier systems and thus, towards networked research information infrastructure or research information space. Each initiative's coverage is slightly different through history. CrossRef is more concentrating on research output although moving towards a more generic coverage, where ORCID and ResearcheID have the researcher in the center connecting to output. Handle originates from the library domain towards a sustainable, accessible, secure and metadata-rich infrastructure as does VIAF, providing building blocks, i.e. authority files. ORCID, ResearcherID, and CrossRef with DOIs clearly address *identification* in their names, however differ from URIs, UUIDs and OpenIDs in being more technology driven, where the former do rather focus on the research entities as such.

5 Summary and Conclusion

From the presented use-cases, the need and thus motivation for globally unique and persistent identification of research entities is clear. The CERIF task group started discussions over naming and conceptual formalization, i.e. modeling in CERIF a Federate Identifier entity. However, the design and concept is not yet entirely clear. Here, we investigated identifier systems and initiatives in the wider scholarly domain to better understand the key issues with identification. Identification is recognized as a significant entity and this contribution will be forwarded to the CERIF task group for continued discussion and uptake in the forthcoming CERIF model release. The idea of globally unique identifiers is exciting, governance and security issues must not be neglected. We support Clifford Lynch, the director of the Coalition for Networked Information (CNI)[18] in that the lack of interconnection is striking (Enserink 2009), but „there's nothing wrong with letting a couple of systems evolve, he says; they can always be linked or merged later on."

[18] Coalition for Networked Information (CNI): http://www.cni.org/

Acknowledgements

We wish to thank the CERIF task group for stimulating and highly fruitful discussions and supportive contributions. The work was partly supported also by EC Funding under META-NET with grant agreement no. 249119.

References

Bizer C., Heath T. and Berners-Lee T. (2009): *Linked data – the story so far*. International Journal on Semantic Web and Information Systems. 5(3) pp. 1-22.

Bischof C., Bunsen G. and Müller J. (2008): Die RWTH-Kundennummer – Eine einfache aber robuste Identitätsnummer, DFN Mitteilungen 72, pp. 19-21, 2007.

CrossRef (2009): A short history: http://www.crossref.org/08downloads/CrossRef10Years.pdf

Enserink M. (2009): *Are You Ready to Become a Number?* Science, pp. 1662-1664.

Fenner M. (2012): *ORCID Unique Identifiers for Authors and Contributors*. Conference Slides *The Value of Unique Scholarly Identifiers to Academics, Institutions and Countries*. February 2012. http://about.orcid.org/content/orcid-unique-identifiers-authors-and-contributors

Evermann, J. and Wand Y. (2001): *Towards Ontologically Based Semantics for UML constructs*. InProceedings: ER2001. Kunii, H.S, Jojodia, S. and Sølvberg, A. (Eds.) pp. 354–367..

van Godtsenhoven K., Karstensen M.E., Sierman B., Bijsterbosch M., Hochstenbach P., Russell R. and Vanderfeesten M. (2009): *Emerging Standards for Enhanced Publications and Repository Technology: Survey on Technology*. Amsterdam Univ. Press, SURF/EU Driver Series.

Heath T. and Bizer C. (2011): *Linked Data - Evolving the Web into a Global Data Space.* (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, Morgan & Claypool, pp. 1–136,.

Hoellrigl T., Dinger J., Hartenstein H. (2010): *FedWare: Middleware Services to Cope with Information Consistency in Federated Identity Management*, Proceedings of the Fifth International Conference on Availability, Reliability and Security (ARES 2010), pp. 228-235, Krakow, Poland, February, 2010.

Hoellrigl T., Schell F., Hartenstein H. (2009): *Federated Identity Management as a Basis for Integrated Information Management*. Information Technology, Vol. 51(1), pp. 14-23, March, 2009.

Hoellrigl T., Schell F., Hartenstein H. (2008): *Federated and Service-oriented Identity Management at a University*, Proceedings of the 14th European University Information Systems (EUNIS 2008), pp 59, Aarhus, Denmark, June, 2008

Jörg B., Ruiz-Rube I., Sicilia M.-A., Dvorak J., Jeffery K., Höllrigl T., Rasmussen H.S., Engfer A. Vestdam T., Garcia Barriocanal E. (2012): *Connecting Closed World Research Information Systems through the Linked Open Data Web*. International Journal of Software Engineering and Knowledge Engineering (IJSEKE), Vol. 22, *Consuming and Producing Linked Data on Real World Applications*. June, 2012.

Jörg B., Ferlez, F., Uszkoreit H., Jermol, M. (2008): *Analyzing European Research Competenties in IST: Results from a European SSA Project.* In Proceedings: CRIS 2008, Maribor, Slovenia.

Kahn R. and Wilensky R. (2006): *A framework for distributed digital object services*. International Journal on Digital Libraries (2006). Vol 6(2), pp. 115—123.

Mylopoulos, J. (1992). *Conceptual Modeling and Telos*. In P. Locuopoulos and R. Zicari, (Eds.) Conceptual Modeling, Databases and Cases. John Wiley & Sons, Inc., New York et. al.

Wand, Y. and Weber, R. (2002): *Research Commentary: Information Systems and Conceptual Modeling – A Research Agenda.* Information Systems Research. Vol. 13(4), pp. 363—376.

Contact Information

Brigitte Jörg

German Research Center for Artificial Intelligence (DFKI GmbH)
Alt-Moabit 91c
10559 Berlin
Germany

brigitte.joerg@dfki.de

Thorsten Höllrigl

AVEDAS AG
Waldstrasse 65
76133 Karlsruhe
Germany

T.Hoellrigl@avedas.com

Miguel-Angel Sicilia

University of Alcalá
Plaza de San Diego s/n
28801 Alcalá de Henares
Spain

msicilia@uah.es