ODaF Europe 2009

MEETING REPORT

International Data Service Center
Institute for the Study of Labor
Bonn, Germany
April 2nd-3rd 2009

Overview

The ODaF Europe 2009 regional meeting was held this year at the International Data Service Center (IDSC) of the Institute for the Study of Labor (IZA) in Bonn, Germany on April 2-3. The focus was on metadata in research data centers. We would like to express thanks to IZA's International Data Service Center (IDSC) for its support and assistance in making this event possible.

The event was attended by over 20 participants from academic, research, archives, funding and international agencies, with a strong representation from Germany and the Netherlands. After going through a brief introduction to ODaF and an update on our completed, ongoing and planned activities, day one consisted in thematic presentations by the participants (see agenda below).

These high quality presentations covered a broad range of subjects and provided ideas and topics for discussion for the second day of the meeting. This supported a series of question/answer sessions and debates that translated into a list of prospective projects or activities for participating agencies and ODaF.

Agenda

Day 1

8:30-9:00	Registration / coffee / breakfast
9:00-10:30	▪ Welcome ▪ Participants' Introduction ▪ ODaF Overview and Activities [Arofan Gregory/Pascal Heus, ODaF] ▪ Metadata in Research Data Centers [Arofan Gregory/Pascal Heus, ODaF]
	Break
11:00-12:30	▪ The IDSC of IZA: Past, Present and Future [Nikos Askitas, IZA] ▪ Concept Hierarchy Assisted News in Economics [Georgios Tassoukis, IZA) ▪ The European Values Study and its Metadata: Stock Taking and the Future [Ruud Luijkx, U. Tilburg] ▪ The DRIVER Initiative for Networking Repositories [Wolfram Horstman, DRIVER]
	Lunch break
13:30-15:00	▪ Enhanced Publications [Thomas Place, U. Tilburg] ▪ The DatapluS Enhanced Publications Editor [Bart van Nieuwburg, CenERData] ▪ Data Documentation and Dissemination with Questasy [Alerk Amin, CenERData] ▪ Open Access and Research Data [John Doove, SURF]
	Break
15:30-17:00	▪ Overview, Importance and Benefits of SDMX Registries [Arofan Gregory, ODaF] ▪ SDMX at the European Central Bank[Gerard Salou / Xavier Sosnovsky, ECB] ▪ Virtual Research and Collaborative Center [Pascal Heus, ODaF/NORC] ▪ Topics for discussion on day 2
19h00 - 22h00	IZA Dinner at Restaurant Schänzchen

Day 2

8:30-9:00	Coffee / breakfast
9:00-12:30	Follow-up discussions, ideas, brainstorming based on day 1: ▪ Metadata: ingestion, archival, digital resources, project, researcher ▪ Enhanced Publications ▪ Standard mappings ▪ Any other idea, topics, suggestion raised by participants

	A coffee break will be taken at 10h30
12:30-13:30	*Lunch*
13:30-16:00	▪ Developments of the Data Infrastructure in Germany since the end of the 90s [Hilmar Schneider, IZA]
14:00-16:00	▪ Action plan, next steps, follow-up, conclusions

Discussion Topics

One of our aims of organizing ODaF meetings in Europe and North America is to stimulate the community involvement with ODaF. The second day of the meeting is therefore used to make an inventory of ideas and interests among people, institutions, consortia, etc. for which ODaF involvement could bring added value. The topics listed below were raised and discussed during the 2nd day of the meeting at IZA. ODaF will follow up with the participants who were at the meeting and inform all other ODaF members, whether these topics should be a target for a focused ODaF activity, paper or project.

Persistent identifiers

The challenge of being able to globally, uniquely and persistently identify an resource or object remain an issue that seem above the level of a single standard (DOI, DDI, SDMX, etc.). The question of "how to do it" seems more an organizational challenge than a technical one. The Open Data Foundation could potentially publish recommendations on this issue or engage with OASIS (or similar organizations such as the W3C or ISO) to provide a discussion space and establish guidelines. The initial necessary step however to accomplish this would be to more precisely document what are the issues at hand.

This effort will be coordinated by Rob Grim who will:
- Help compile / aggregate existing information
- Identify a set of individual and agencies interested in the project
- Advertise the topic amongst the community and call for feedback

The outcomes will be made as a report or white paper made available through ODaF who will facilitate the process. This will be used to drive potential next steps.

ODaF Coordinator: Rob Grim (First draft to be delivered in June 2009).

Formalization/publication of concept hierarchies (DDI/SDMX/ISO/IEC 11179)

Issues with managing/maintaining concept hierarchies

A wide range of common metadata collections are used on a daily basis by producers, archives and researchers for various purposes. A few examples (amongst many) include classifications such as the ISO country codes, national employment or education codes (ISCED), or concepts lists such as the Integrated Metadata Base (IMDB) in Canada or HASSET in Europe. These metadata collections are typically available on the web as HTML web pages, PDF documents and in some cases Excel spreadsheets. While good for human consumption, these formats are not particularly fit for consumption by applications that need to understand the structure or meaning of the information.

Metadata specifications such as DDI, SDMX or ISO 11179 provide a model to capture such information in an XML format that can essentially be published as documents or through the web. Capturing and representing code - and concepts lists in a structured format will enhance the discovery, accessibility and reuse of metadata and greatly facilitate the use and processing of data by applications. Maintenance of the code lists is the responsibility of the agency in charge of the collection (i.e. ISO for the country codes, BLS for the employment codes in the USA, etc.). ODaF could play a significant role however in providing and disseminating machine actionable instances for the frequently used code and concept lists in social science research.

The Open Data Foundation and other interested partner agencies and individuals can collaborate to convert the frequently used standard code and concepts list into machine actionable formats. Access to this (structured) content could be provided through ODaF or through the organization that maintains the code and concepts lists. This initiative would facilitate access to common metadata in an XML format and foster the adoption of specifications (i.e. DDI 3.0, SDMX) by maintenance agencies. Any issues around intellectual property or licensing will be identified in collaboration with the maintenance authorities.

Project activities would include:
- Identifying an initial set of collections to be captured
- Contacting official maintenance agencies to inform them about the initiative and probe for interest in publication/maintenance on their side
- Developing simple tools for conversion of metadata to XML
- Specifying and deploying an infrastructure for publication of code and concepts lists on ODaF server
- Defining strategy for maintenance / sustainability and potential expansion
- After deployment:
    - Informing community of availability metadata
    - Examining potential expansion of common metadata collections
    - Monitoring of system usage

The initial team to seed this project proposal is composed of Arofan Gregory (ODaF), Pascal Heus (ODaF), and Georgios Tassoukis (IZA). Chris Nelson (Metadata Technology) and the CESSDA PPP[1] Work Package 1 & 3 members have been mentioned as potentially interested parties. Necessary resources will need to be identified to support the above activities.

ODaF Coordinator: Pascal Heus

HASSET/ELSST: licensing issues, process, etc.

The Humanities and Social Science Electronic Thesaurus or "HASSET" (http://www.data-archive.ac.uk/search/hassetAbout.asp) is a subject thesaurus which has been developed by the UKDA over the past 20 years. Initially based on the UNESCO thesaurus compiled by Jean Aitchison (Paris: UNESCO, 1977)(ISBN: 92-3-101469-2) HASSET has been continuously expanded and updated for use in its online retrieval system. HASSET is also the basis for a multi-lingual thesaurus – the European Language Social Science Thesaurus (ELSST) - developed under the MEDIERA project (http://www.data-archive.ac.uk/randd/madiera.asp).

This is an attractive product that could potentially be adopted by several organizations other than the UKDA or members of CESSDA. Its reuse would also further harmonize metadata through the use of common keywords and could potentially bring new contributions to HASSET (feedback, expansion. domain specific versions, translation into additional languages, etc.).

Using HASSET or ELSST however raise some organizational and legal issues. For the English version, maintained and copyrighted by the UKDA, a procedure is in place for its distribution under a non-commercial use license (basically, a license to sign which says an organization will not pass it on, will not make money out of it and will always accredit the source). ELSST (9 language version) is also maintained by UKDA at present but ownership is not well defined and it will probably be held by the new CESSDA ERI[2] once it becomes a legal entity. Part of the present CESSDA PPP is to produce new maintenance software for ELSST where responsibility for maintenance passes to a CESSDA management team. ELSST was an output from two EU projects LIMBER and MADIERA, hence there is a similar distribution license but has up till now only been used for distribution to CESSDA project members and it is not clear how this could be applied to non-members. In addition, to the licensing issue, it is not clear how the thesaurus can be extended or be used to produce derivative work such as domains specific subsets.

We know of a few agencies that have expressed interest in using HASSET / ELSST (IZA, IHSN, CentERdata, UvT) and therefore we propose to come together under the ODaF to open a discussion

[1] http://www.cessda.org/project/wps.html
[2] European Research Infrastructure

with CESSDA. This will aim at jointly expressing our interest in the thesaurus, advocating for an open licensing model that would make ELSST available to non-CESSDA member agencies, and addressing maintenance and derivation issues.

A small team could be formed to lead this project and perform the following activities:
- Identify and bring together non-CESSDA agencies interested in adoption HASSET / ESSLT
- Examine licensing and related legal issues
- Examine the thesaurus update, extension process, and derivative work issues
- Document the preferred delivery format options (RDF, XML, database, excel, etc.) and publication processes
- Initiate contact with the UKDA and CESSDA to discuss the above issues

The project scope could also be expanded to the identification of other thesauri of interest (OECD, Statistics Canada, UNECE, etc.) and document their availability and potential use.

During the meeting, Nikos Askitas (IZA), Georgios Tassoukis (IZA), and Alerk Amin (CentERdata) have expressed interest in this activity. The International Household Survey Network is also known to be interested in using ESSLT.

ODaF Coordinator: Pascal Heus

Standards SLAs for registry scenarios

Operating metadata registries and maintaining the underlying ICT infrastructure is a task that comes with considerable responsibilities. Legal protection is an important aspect of this that if often forgotten and not well understood. This typically involves establishing Service Level Agreement (SLA) with the various stakeholders: users, data/metadata providers, hosting and other service providers (sub-contractor). Given the importance of registries in modern metadata management framework, documenting these issues, providing guidelines and drafting templates would be helpful to the community.

Putting together such set of information and documents could be driven through an ODaF project whose activities would involve:
- Documenting legal issues involved in operating and maintaining a registry infrastructure
- Documenting use cases
- Develop checklists
- Prepare sample Service License Agreement templates for
    - End Users (paid / unpaid)
    - Data / Metadata Providers
    - Sub-contractors (hosting, backup/disaster recovery, security, connectivity, cloud, etc.)

This project would be coordinated by Arofan Gregory (ODaF) and executed in collaboration with DRIVER with potential inputs from ODaF experts such as Bob Glushko (OASIS, Berkeley's Center for Document Engineering) and Eduardo Gutentag (OASIS) ODaF experts would be consulted for this project.

ODaF Coordinator: Arofan Gregory

Reference architecture for Social Sciences data

Having a model for an IT architecture to support social science data and metadata would be very helpful for the many organizations interested in adopting standards and best practices. Several initiatives are aiming in this direction and a coordinated effort under the ODaF framework would facilitate bringing together ideas and harmonize guidelines.

Such project could aim at:
- Identify and examine key issues i.e. use of harvesting vs registry architectures, technology implementations applications that make use of virtualization, grid and cloud computing
- Policy issues, i.e. what level of content aggregation is preferable, achievable? Discipline or domain specific at a national or international level? ,

- Identifying major players and interested organizations around the globe such as the National Science Foundation, Library of Congress, Science Commons in the US or University of Manchester and others in Europe.
- Produce general recommendations on various approaches, tools and techniques for the implementation of an harmonized architecture

Turning this into a project proposal will require some additional research and Rob Grim will coordinate this initial effort.

ODaF Coordinator: Rob Grim
Draft target date: December, 2009

Metadata Standards Crosswalks

ODaF, through various initiatives such as METIS, is already involved in developing mappings between metadata standards such as DDI, SDMX, ISO 11179, ISO 19115 and Dublin Core.

More effort is also needed to support the workflow of digital objects management with Fedora, e-Prints and DSpace for institutional repositories, data archives and data centers. Furthermore (granular) data discovery services need to be improved as well as content aggregation and dissemination services. Summarizing the work that has already been done and identifying use cases that illustrate how *between standards mappings* can support workflows for storage, ingest, discovery, dissemination and retrieval of (meta)data should be a first part of this effort. The use cases will then act as a 'natural' starting point to map the fields of the vocabularies that are at stake. Additional tools probably need to be developed (along with recommendations on how to implement these) to support these objectives.

Rob Grim will coordinate this initial assessment and aim to bring together a team of experts to turn this into a project proposal. This initiative might be aligned with a proposal that addresses both social science infrastructural issues and the provision and dissemination of semantically rich content (i.e. enhanced publications).

ODaF Coordinator: Rob Grim. (First draft to be delivered, November 2009).

Remote Computing with JoSuA

The IDSC at IZA has developed an instrument for controlled remote data processing known as JoSuA (Job Submission Application). Originally designed in order to grant international researchers access to German labor market data, JoSuA has matured into a flexible instrument of data analysis of configurable degree of automation designed to fit the needs and specifications of each individual data provider.

JoSuA is therefore suitable for use by data providers who own such data and wish to make it available to a larger research community without jeopardizing the security of their data.

Data providers may:
• own JoSuA
• use JoSuA "as a service" with data hosting at the IDSC Data Enclave
• use JoSuA "as a service" without data hosting at the IDSC Data Enclave

In all three flavors, JoSuA gives data providers full control of the output's censoring. JoSuA is at the heart of several IDSC partnerships with agencies such as the RDC of IAB, the RDC of IQB, ROA and others.

IZA is considering making JoSua available to other organizations and would like to know if there is enough interest to justify for the tool to grow beyond the prototype stage. Further, they would like to have a large installed base for further testing and evaluation.

We believe ODaF can support such effort by promoting the tool amongst the community and establish connection with potential adopters. This is more an activity than a project and we will remain in contact with IZA (Nikos) on this topic.

ODaF Coordinator: Pascal Heus

Web 2.0 in social science

Taking advantage of emerging web technologies often nicknamed “Web 2.0” seems to be on everyone’s agenda. This topic was already raised at past ODaF meeting in the US and Europe and there is a clear demand for such products. How this technology can be used effectively for social science metadata or knowledge capture however is not clear and putting together a research group would help answer such question. This can cover a broad range of topics such as:

- Requirements for metadata visualization (EVS)
- Integration in various tools
- Use of microdata with services such as ManyEyes, Swivel, GapMinder and others
- Implement and provide leverage on social tagging / folksonomies
- Synergies with the semantic Web and RDF (express XML metadata in RDF)
- Etc.

The output of this research effort would be a “Findings” paper answer question such as:

- What are the various domains of application / use cases?
- Where does it work /does not work
- Which tools are available today (showcase)?
- UI mining / UI for editing
- Identify DDI Dimensions

Several participants expressed their interest in this (Goergios, Rob, Jenny, Alerk, Rene, Christian) and Pascal will coordinate the initial effort as well as establish connections with other individuals / agencies. The aim will be to setup a web 2.0 research project whose first task will be to identify the sub-topics of interest and outline the content of a Findings document. The project output could be virtual and maintainable over the web.

ODaF Coordinator: Pascal Heus

Meta-language for stats packages

Having the ability to capture statistical expressions in a non-proprietary syntax would be very helpful for preservation and generic processing purposes. Covering all possible operations is a daunting challenge and focusing on specific subsets seems to be a more realistic approach to this problem faced by several initiatives and organization (DDI, DANS, etc.).

This issue will be further examined by a small team composed of René, Nikos, Jenny, Pascal and possibly other contributors that will focus on basic data transformations and data disclosure operations (essentially reshaping operations) to produce a paper on the topic. This activity output could be used as a basis for implementation or integration in other packages or XML specifications.

ODaF Coordinator: Pascal Heus

Translation and standards – what to do?

When working with multi-lingual metadata or data, agencies often use specialized software to perform the translation operations. These packages often have the ability to import / export text in various formats including XML. Having the ability to easily exchange text contained in DDI, SDMX or other specifications commonly used in social sciences with such packages will illustrate the practical usefulness of metadata reuse and will stimulate the adoption of the standards across a range of domains and communities.

We propose to put together a team composed of Alerk, Jenny, Rob, Pascal and possibly other contributors to produce a paper on this topic. This will identify commonly used packages, describe their import/export formats, and when possible provide mappings between the DDI and other standards. This could then be used as a basis for the implementation of tools. A starting point for this group will be to contact Karl at ICPSR who wrote a paper on this topic.

ODaF Coordinator: Pascal Heus

Infrastructure/enhanced publications

The meeting at IZA provided a space where representatives of two separate worlds could meet: a world where the provision and exchange of structured statistical content is core (SDMX), and a world where unstructured text corpora and other digital resources used in humanities research are dominant (Driver). An interesting discussion emerged on the preferred infrastructure requirements for large scale data dissemination and intensive data exchange services. One idea that emerged from the discussion was to set up a joint ODaF/Driver grass roots environment that enables us to test i.e. embedded data services that could be part of an enhanced publication and access to distributed (data) resources that are used for (large scale) content aggregation.

ODaF will seek collaboration with Wolfram Horstman (Driver) to set up the environment and to articulate the requirements for effectively testing the use cases outlined above. A joint ODaF Driver project proposal might be submitted for funding to allocate necessary resources.
ODaF: Rob Grim (Coordinator), Arofan Gregory and Pascal Heus.
Draft target date: November 2009.

IZA Red Cube seminars?

The IDSC at IZA recently started to organized seminars[3] to provide a forum for high-quality technology presentations related to the institute's research context. With international lecturers and an audience from Germany and neighboring European countries, it aims at becoming a focal point for data technologists and data analysts. Presentations of broad interest will take place biweekly and are open to the public, while the internal working seminars are targeted at members of the IZA technology group and may have a more narrow and application-oriented focus.

Many of the presentations are addressing topics of interest to the social science metadata and technology communities and it would be highly beneficial to have the ability to open access to a global audience. We discussed some of the technology options such as having the ability for virtual participation (i.e using Adobe Connect) or providing replays over the web (YouTube, Google Video or other streaming server). Entry level video capture equipment can be used for such events (Flip, Kodak Zi6) and simple software like Camtasia Studio or Adobe Captivate can be used to capture the presentations.

ODaF would gladly provide its technical input on how the Red Cube seminars could be made available to a broader virtual audience and welcome the opportunity to advertise the events.

ODaF Coordinator: Pascal Heus

Merging Micro and Macro Data Sources (DDI – SDMX)

One idea which has been proposed comes from the availability of many sources of public, official aggregate data from members of the SDMX community. Data now available, or soon to be made available, in standard SDMX form includes:

- All public data from the OECD
- Significant data from the Federal Reserve Board (industrial production, etc.) and the Federal Reserve Bank of New York (exchange rates, market interventions)
- All of the European Central Bank public data
- Other central banking data (through the Bank for International Settlements)
- Education data (through Eurostat and UNESCO)
- National Accounts data (and some other) from the IMF

This data could be of significant interest to many researchers, who would wish to combine it with micro-data in their analysis. The idea would be for ODaF to host an SDMX registry providing a single place for potential users to be able to search for this data. ODaF could partner with other organizations or

3 http://www.iza.org/index_html?lang=en&mainframe=http%3A//www.iza.org/en/webcontent/events/idsc_red_cube_seminar&topSelect=events&subSelect=seminar

projects which might want to use these data sources, and the search and updating facilities available through the ODaF registry. This cross-walk could potentially be based on a mapping of SDMX and DDI.

The registry would require that a mechanism be built to regularly query the available data sets and update the registry, since the data providers in question would probably not be interested in registering the data themselves.

There is also interest from Statistics Sweden and the Australian Bureau of Statistics to provide hooks into Geographical Information Systems.

Many of those interested in this project were not attending the ODaF Europe 2009 meeting, so this topic will need to be further investigated. This idea will be pursued by Rob Grim and Arofan Gregory.
Draft target date: October 2009.

How to setup a Research Data Center

From the technological and organizational perspectives, establishing and maintaining a research data center (RCD) or data enclave (DE) is a challenging enterprise. Given the number of RDC, data, metadata and ICT experts we have available that have been involved in such process, putting together our experiences and recommendations would provide valuable knowledge to other organizations running and planning to launch an RDC. These could be compiled in an ODaF white paper that:

- Document what to think about
- Establish high level checklists
- Examine various models: Stand Alone (IZA), Networked (Canada RDC), Virtual (NORC), etc.
- Provide case studies
- Share lessons learned
- Etc.

The initial step would be to confirm general interest in such document and identify potential contributors.

ODaF Coordinator: Pascal Heus

Other issues

The following topics were raised but were not discussed or did not result in specific actions:

- Keywords vs. topics vs. concepts
- SURF Legal issues – anything ODaF can do to help?

Next Steps

Projects/Activities

As initially mentioned, the list of projects and activities listed above are prospective and the initial follow-up actions will be to:

- Confirm interest by listed individuals / agencies and circulate these ideas amongst ODaF community
- Refine the scope & timeline
- Estimate and secure resources (in kind / funding)
- When applicable, turn into a formal project/activity for approval by directors

Plan for ODaF Europe 2010

Rob Grim (UvT) and René van Horik (DANS) briefly discussed the idea of jointly organizing the 3rd ODaF Europe meeting 2010 in the Netherlands. Rob will contact René to find out what needs to be done to confirm these plans and which further steps need to be taken. This is not to say however that other suggestions or initiatives for organizing the next ODaF Europe meeting are to be discouraged.

Note that we haven been asked if possible to have the ODaF 2010 meeting to be scheduled around the UK Royal Statistical Society (RSS) metadata meeting that will be organized jointly by the RSS Statistical Computing section and the Association for Survey Computing (ASC). This venue could also be an option for the ODaF 2010 meeting.

Support / Collaboration

For ODaF to grow beyond a team of motivated individuals into a self-sustained organization, we need to expand the list of active contributors and establish better funding strategies. We will continue to consult with our members on this topic and would welcome suggestions. One of our current goals is to grow into an agency of 15-20 active individuals in the next 3-5 years.

List of Participants

Thomas Leopold	Bamberg University	thomas.leopold@uni-bamberg.de
Marcel Raab(*)	Bamberg University	Marcel.Raab@uni-bamberg.de
Christian Gerhards	Bielefeld University	christian.gerhards@uni-bielefeld.de
Stefan Liebig	Bielefeld University	stefan.liebig@uni-bielefeld.de
Alerk Amin	CenERData	a.amin@uvt.nl
Bart van Nieuwburg	CenERData	b.vannieuwburg@uvt.nl
Jannik Jensen	Danish Data Archive	jvj@dda.dk
Rene van Horik	DANS	Rene.van.Horik@dans.knaw.nl
Wolfram Horstman	DRIVER	wolfram.horstmann@uni-bielefeld.de
Markus Quandt	GESIS	markus.quandt@gesis.org
Uwe Jensen	GESIS	uwe.jensen@gesis.org
Gerard Salou(*)	European Central Bank	Gerard.Salou@ecb.int
Xavier Sosnovsky	European Central Bank	Xavier.Sosnovsky@ecb.int
Denis Huschka	RatSWD	dhuschka@ratswd.de
Ruud Luijkx	Tilburg University	R.Luijkx@uvt.nl
Thomas Place	Tilburg University	T.W.Place@uvt.nl
Andreas Aschenbrenner	SUB Goettingen	aschenbrenner@sub.uni-goettingen.de
Stefanie Ruehle(*)	SUB Goettingen	sruehle@sub.uni-goettingen.de
Jenny Linnerud	Statistics Norway	jal@ssb.no
John Doove	SURF	Doove@surf.nl
Nikos Askitas	IZA	askitas@iza.org
Georgios Tassoukis	IZA	tassoukis@iza.org
Arofan Gregory	ODaF	agregory@opendatafoundation.org
Pascal Heus	ODaF	pheus@opendatafoundation.org
Rob Grim	ODaF	rgrim@opendatafoundation.org

(*) Could not attend