ADOPTING RDA OUTPUTS FOR...
DIGITAL LANGUAGE RESOURCES

RESEARCH DATA ALLIANCE

CLARIN adopts 3 RDA recommendations for digital language resources management

CLARIN is a distributed infrastructure of centres providing access to digital language data (in written, spoken, video or multimodal form) and tools to discover, explore, exploit, annotate, analyse or process this data, independently of where the data is located. CLARIN relies on a well-defined architecture of repositories, providing access to the language data sets, which are documented via metadata and citable via persistent identifiers. To communicate – with both insiders and outsiders – a lingua franca is essential to ensure all parties are talking about the same thing for instance when they refer to a digital object, a bitstream or digital collections. The RDA's Data Foundation and Terminology provides this lingua franca. Using its recommendations as a common ground also greatly enhances the work being done in other RDA working groups. Having adopted the Data Foundation and Terminology, CLARIN was facilitated in using also other RDA recommendations, such as Metadata Standards Directory and Repository Audit & Certification Catalogues and fully exploiting their combined benefits.

The Challenge

"Human language often comes with ambiguity. Communicating about research data management is no exception to that: it happens often that two experts are referring to different concepts with the same word. It also happens the other way around – multiple words referring to the same thing. In order to streamline the communication and to avoid needless and time-consuming interactions on terminology, the use of a common and well-defined vocabulary is strongly advised."

Says Dieter Van Uytvanck, Technical Director at CLARIN, the adopting organisation.

"Adoption of the vocabulary and associated documentation provided by the RDA Data Foundation and Terminology working group came natural to CLARIN as it fits very well with the way CLARIN's infrastructure is setup": a networked federation of distributed language data repositories, service centres and knowledge centres. Through CLARIN, tools and data, accessible from different centres, are interoperable, so that data collections can be combined and tools from different sources can be chained to perform complex operations to support researchers in their work. Continues Van Uytvanck: "Without a terminological common understanding we would – in the best case – loose a lot of time in repeated semantic discussions or – worse – end up with conflicting interpretations about our data architecture". Furthermore, using the RDA Metadata Standards Directory and Repository Audit and Certification Catalogues enables researchers to find easily the diverse contents accessible through the CLARIN infrastructure ensuring also the reliability of sources and facilitating researchers' analysis, correlation and processing of language data.

RDA RECOMMENDATIONS ADOPTED

Data Foundation and Terminology: simplifies understanding and communication about basic concepts such as digital object and persistent identifiers.

Metadata Standards Directory: enables discovery of metadata standards for documenting research data, regardless of academic disciplines, and addresses issues related to coverage, ease of maintenance and sustainability.

Repository Audit and Certification Catalogues: creates harmonized common procedures for certification of repositories at the basic level, drawing from the procedures already put in place by the Data Seal of Approval (DSA) and the ICSU World Data System (ICSUWDS).

ANSWERING COMMUNITY NEEDS

CLARIN is for all Humanities and Social Science disciplines, as far as they work with digital language resources (text, multimedia, lexical information etc.). Although the focus of CLARIN activities may vary from country to country, and although many of those who participate in developing CLARIN have a background in linguistics, language technology or computer science, CLARIN is broadly used by scholars of many disciplines, for instance literature studies, history, linguistics, sociology, psychology, computational linguistics, philosophy, and ethnology.
Furthermore, the CLARIN user community is spread all over Europe and by extension all over the world, resulting in an extremely heterogeneous set of requirements.

WHY RDA

CLARIN is an active community within the RDA contributing to a number of interest and working groups relevant to the field of language data and its research activities. This has been very beneficial, both for CLARIN and for RDA. Feedback about linguistics community-specific issues were smoothly integrated into the RDA working groups outcomes: CLARIN's adoption ensured that an early and continuous validation with the community practices was always in place.

Find out more

Visit RDA @ rd-alliance.org
Email: enquiries@rd-alliance.org

The Adoption

CLARIN gradually introduced the DFT concepts in defining documents and day-to-day communication. This laid the basis for adopting also the Metadata Standards directory, and the Repository Audit & Certification Catalogues RDA recommendations. Using RDA's recommendations supported the establishment of unambiguous and clear communication about language research data management across the distributed centres hosting the CLARIN data and services.

CLARIN gradually introduced the DFT concepts in defining documents and day-to-day communication. This laid the basis for adopting also the Metadata Standards directory, and the Repository Audit & Certification Catalogues RDA recommendations. Using RDA's recommendations supported the establishment of unambiguous and clear communication about language research data management across the distributed centres hosting the CLARIN data and services.

The Data Foundation and Terminology Working Group has been on CLARIN's radar from an early stage. By contributing to it and participating in the working group discussions the CLARIN team was able to appreciate the benefits of becoming an (early) adopter and using RDA's recommendations as the foundation for a systematic approach to managing language digital data offering researchers a rich set of interoperable, distributed data and services managed independently.

Lesson Learnt

- **Even if you do not foresee it, many other communities are confronted with similar issues related to research data management and processing. Discussing these problems – and possible solutions– can save a lot of hassle.**
- **Being involved in the RDA work from the start helps a lot in ensuring that the community practice is reflected and recognized in the recommendations. At the same time it stimulates looking at solutions from another perspective.**

Name: Van Uytvanck
Surname: Dieter
Qualification: Technical Director
Affiliation: CLARIN ERIC
Country: NL
Email: dieter@clarin.eu
Phone: +31-(0)850091363

CLARIN Common Language Resources and Technology Infrastructure

CLARIN (Common Language Resources and Technology Infrastructure) makes digital language resources available to scholars, researchers, students and citizen-scientists from all disciplines, especially in the humanities and social sciences, through single sign-on access. CLARIN offers long-term solutions and technology services for deploying, connecting, analysing and sustaining digital language data and tools. CLARIN supports scholars who want to engage in cutting edge data-driven research, contributing to a truly multilingual European Research Area.