Outputs from the RDA Linguistics Data Interest Group

RDA 14th Plenary, Helsinki, Finland
Adoption and Outputs Plenary Session
25 October 2019

Andrea Berez-Kroeker, Lauren Gawne, Helene N. Andreassen, Philipp Conzett
Our perspective and engagement with RDA

IG endorsed in 2017.

Co-chaired by Helene N. Andreassen (UiT The Arctic University of Norway), Andrea L. Berez-Kroeker (University of Hawai‘i at Manoa), Lauren Gawne (La Trobe University).

Around 100 members.
Our perspective and engagement with RDA

Main objectives:

Development and adoption of common principles and guidelines for data citation and attribution by professional organisations, academic publishers, and archives for linguistic and language data.

Education and outreach efforts to make linguists more aware of the principles of reproducible research and the value of data creation methodology, curation, management, sharing, citation and attribution.

Efforts to ensure greater attribution of linguistic data set preparation within the linguistics profession.
Output 1: Survey of data citation practices in linguistics

A survey of data citation practices in 9 linguistics journals over a ten-year period shows that we do not have a culture of citing data in our publications:


Reproducible research in linguistics: A position statement on data citation and attribution in our field. Linguistics 56(1): 1–18. doi.org/10.1515/ling-2017-0032 (Currently the most downloaded article in Linguistics).
Output 2: Austin Principles of Data Citation in Linguistics

Published in 2018.

Set of guidelines that enable linguists to make informed decisions regarding the accessibility and transparency of their research data.

For all types of data, all subfields of linguistics.

Based on the FORCE11 Joint Declaration of Data Citation Principles.

www.linguisticsdatacitation.org

Endorsed by 10 organizations and 97 individual linguists.
Output 3: MIT Open Handbook of Linguistic Data Management

Intended audience is linguistics practitioners at all levels (students, researchers, instructors) and in most subfields of linguistics.

Content: 1) Thirteen full-length chapters on conceptual foundations of data management for linguistics and best practices. 2) Around 50 short Data Management Use Cases.

Edited by Andrea L. Berez-Kroeker, Lauren B. Collister, Bradley McDonnell & Eve Koller. More than 70 contributors.
Output 4: Recommendations for Data Citation in Linguistics

Guidelines for data citations in-text and in reference lists, including commented examples and definitions.

Intended audience: editors of scientific journals, researchers who plan to cite or archive data, research data repositories.

To be submitted for RDA community review by the end of the year.
Thank you!
CTS adoption piece at ADP

CESSDA SERVICE PROVIDER FOR SLOVENIA

Janez Štebe, RDA Adoption and Outputs session on Friday, 14:00-15:30, RDA Plenary Helsinki 24. 10. 2019
Slovenian Social Science Data Archives (ADP-Arhiv Družboslovnih Podatkov)

• Founded in 1997
• Slovenian national research data centre for social sciences
  • 600 social science studies data accessible in a data catalogue + 150 metadata only
  • Mainly survey data (from 1960's on), few qualitative, social networks and social media
• member of CESSDA ERIC
• obtained CoreTrustSeal in beginning of 2018
• involved in EU, CESSDA and national projects

ADP Social Science Data Archives
http://www.adp.fdv.uni-lj.si/eng/CTS Certification 2017-2019
Experiences to share in acquiring the CTS

• The archive has been operational for 15 years: expertise gained through collaboration and involvement in CESSDA activities

• Role model: ICPSR, UK DA, DANS, → but then adapted to the size and specificity of the setting: here FSD model helped

• Institutionalisation of organisational setting: few articles introduced in the Rules of the Faculty of Social Sciences, University of Ljubljana (Host organisation of a national data service ADP)

• Key: Sustainability guarantee through institutional setting, financial long-term stability of funding, and national membership in CESSDA ERIC infrastructure.

• Key: learn to speak OAIS language: pre-SIP, SIP, AIP, DIP

• Following the designated community definition in description of processes

• Keeping and regularly updating the written workflow data processing steps

• It’s more about the processes and roles than about technology: after first submission the only comment was that we need to nominate who are holding the roles: https://www.adp.fdv.uni-lj.si/kontakt/
Activities to support CTS in the community

- Visits: A VISIT FROM LITHUANIA:  http://www.adp.fdv.uni-lj.si/blog/2019/blog/a-visit-from-lithuania/


- RDA Node Slovenia joint conference on 14th Novembre 2019:  https://www.adp.fdv.uni-lj.si/dogodki/konferenca-odprti-raziskovalni-podatki/

- Most important:
The CESSDA Trust Group offers both existing and aspiring service providers guidance and support in meeting a range of issues and standards relating to trusted data and services. CESSDA requires the adoption of specific criteria such as the internal obligations required from all members (CESSDA statutes) and the trustworthy digital repository (TDR) requirements set by the CoreTrustSeal.

These goals must be met within the evolving infrastructure (skills, services and technology) of European and international research data science.

CESSDA Trust Group consists of a core of service providers with experience in trust standards and certification and key contacts representing each of the CESSDA members and aspiring members.

The group's goals are met through:

- Guidance, engagement and support to members in understanding, acquiring and maintaining compliance with CESSDA obligations and the requirements of the CoreTrustSeal.
- Monitoring and reviewing compliance at an individual and organisational maturity levels. Engaging with trust-related elements of the CESSDA work plan including other working groups and projects.
- Maintaining an overview of the trust landscape including certification standards and the emergence of the FAIR data principles and the requirements of the European Open Science Cloud (EOSC).
Thank you!

University of Ljubljana
Faculty of Social Sciences
Social Science Data Archive
Kardeljeva ploščad 5
1000 Ljubljana
Slovenia

www.adp.fdv.uni-lj.si
arhiv.podatkov@fdv.uni-lj.si
Arhiv.Druzboslovnih.Podatkov
@ArhivPodatkov
The Core Trust Seal: An adoption story
RDAPlenary, Oct 2019, Helsinki

Dr. Natalie Harrower
Director, Digital Repository of Ireland
@natalieharrower | @dri_ireland
Core Trust Seal: certification...with added benefits!
What is the Digital Repository of Ireland?

- Core Trust Seal certified (2018) trustworthy digital repository
- National repository for archiving, curation, preservation
- Humanities & Social Sciences research data
- Cultural heritage digital collections
- Actively supporting Open Science, FAIR data
- Partner in RDA_EU 4.0; EOSC-FAIR WG, EC FAIR data expert group, OECD Business Models for Sustainable Research Data Repositories
What is the Core Trust Seal?

• A core certification scheme for Trustworthy Data Repositories
• 16 requirements that are mandatory, stand-alone, and equally weighted (metadata, preservation, sustainability, licences, expertise, discovery, citation, re-use, software & hardware…)
• Self-certification, then assessed by Assembly of reviewers
• First step in a global framework for repository certification which includes the extended level certification (Nestor-Seal DIN 31644) and the formal level certification (ISO 16363)
• Legal entity under Dutch law; not-for-profit
Why is the Core Trust Seal important?

• Provides an entry level step to certifying trustworthy digital repositories against agreed common criteria
• Rapidly emerging community consensus that CTS is the first step to choose
• Helps to support FAIR data, RDM and Open Science
• Depositors can trust that their data is safe, well structured and accessible for the long term
HOW STANDARDS PROLIFERATE:

(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)

SITUATION:
THERE ARE 14 COMPETING STANDARDS.

14?! RIDICULOUS!
WE NEED TO DEVELOP
ONE UNIVERSAL STANDARD
THAT COVERS EVERYONE'S
USE CASES. YEAH!

SOON:

SITUATION:
THERE ARE 15 COMPETING STANDARDS.

Source: https://xkcd.com/927/ (CC BY-NC 2.5)
Importance: built by the research & data communities

Core Trust Seal emerged from two previous repository certifications

- RDA-WDS Group on Repository Audit and Certification to create harmonised catalogue
- Merger of two existing schema: DSA (Humanities & Social Sciences) and WDS (Earth & Space Sciences)
- Domain agnostic: interdisciplinary
Importance: Adoption

Since 2017
68 CTS certifications
14 since April 2019
Importance: supported by larger networks

- CTS Noted as an exemplar in the report of the European Commission’s Expert Group on FAIR Data
- Falls under the EOSC umbrella
- Adoption growing globally
- Supported by the umbrella structures and networks of Open Science (top-down)
How does CTS enable FAIR data?

- There are several concepts essential to enabling FAIR data that are implied by, if not explicitly noted in the FAIR principles: one is the need for long-term data stewardship and preservation.
- To be realised, FAIR data relies on an ecosystem of components: policies, DMPs, PIDs, standards, repositories. These components must support FAIR.
- CTS fundamentally about LTP (long term preservation), but goes further.
- Specific criteria for repositories under CTS lead clearly in the direction called for by the FAIR principles.
Research Data Lifecycle
Wider benefits of CTS: reports from adopters

- Valuable process for organisation-wide exchange, discovery, communications (audit of full lifecycle of repository data pipeline).
- Excellent way to review policies, decide what needs better articulation, what could be made public. e.g. DRI monthly review calendar. Documents in place for future funding apps.
- Strengthens funding applications; vehicle to explain business needs higher up
- Creates trust, confidence in the depositing community: builds collections
- Increases the visibility of research data
- Increases professionalism/unique value (National Geosciences Data Centre)
- Uncovers implicit knowledge of processes that may lack proper documentation
- Increases the focus of funders on FAIR data management principles (Protein Data Bank)