Information Integration for Machine Actionable Data Management Plans

Tomasz Miksa

Vienna University of Technology

miksa@ifs.tuwien.ac.at

Agenda

- Introduction
- State of the art
- Mapping
- Requirements
- Limitations
- Data Model
- Conclusion and future work

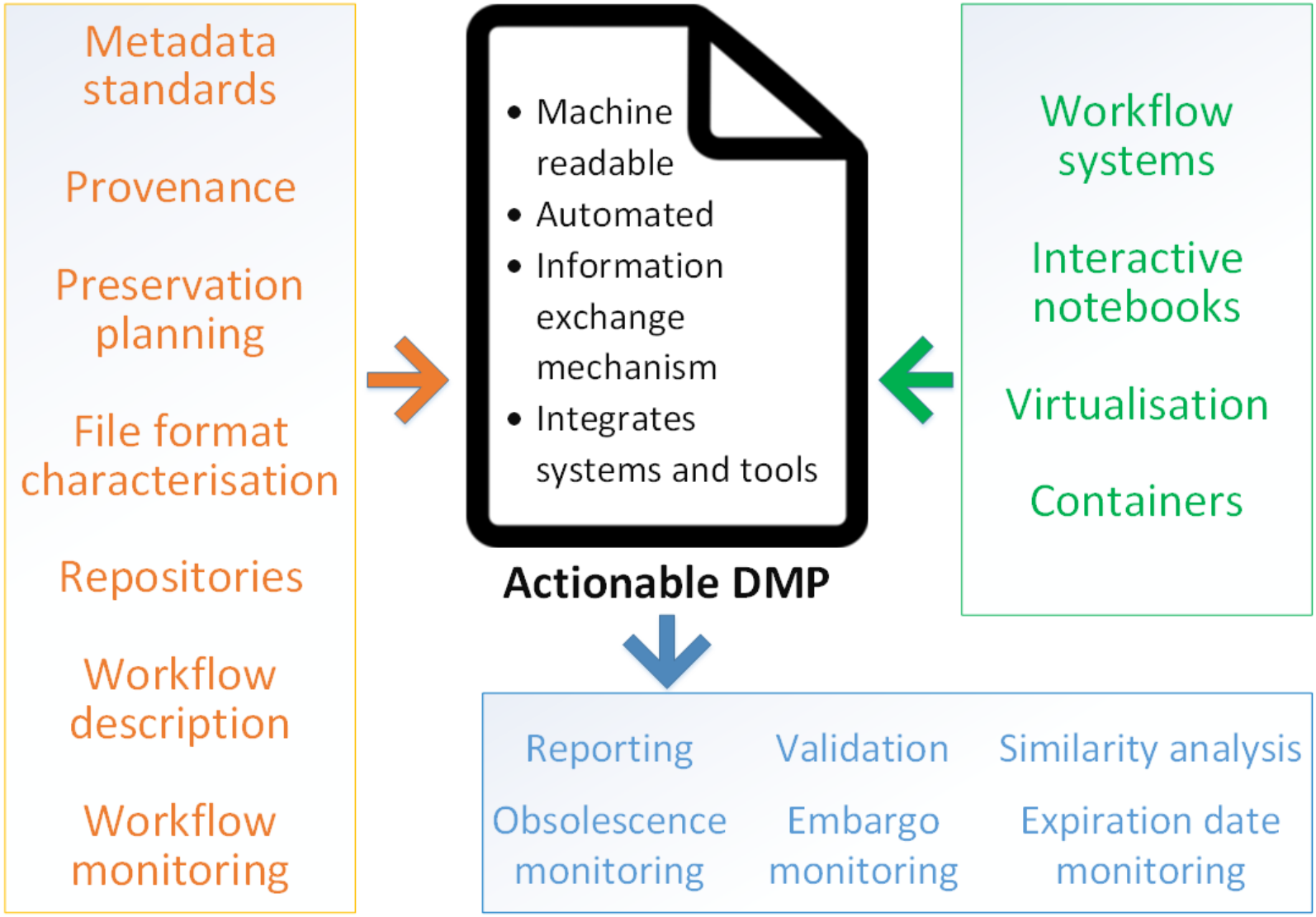
Introduction

- DMPs
  - manually created text documents
  - considered as bureaucracy
  - created too late
  - vague
  - depend on human factor (scrupulousness and awareness)

State of the art

- CERN workshop on Active DMPs (May 2016)
- Research Data Alliance (8th Plenary in September 2016)
  - IG Active DMPs
  - IG Reproducibility
  - IG Preservation e-Infrastructure
- Pending integrations
  - DMP Tool and DMP Online
  - Zenodo and GitHub
  - OSF and DOI
- Recommendations
  - European Open Science Cloud
  - FAIR principles

Hypothesis


Metadata standards
Provenance
Preservation planning
File format characterisation
Repositories
Workflow description
Workflow monitoring
Machine readable
Automated
Information exchange mechanism
Integrates systems and tools
Actionable DMP
Workflow systems
Interactive notebooks
Virtualisation
Containers
Reporting
Validation
Similarity analysis
Obsolescence monitoring
Embargo monitoring
Expiration date monitoring

Mapping

Dublin Core, PREMIS,...

W3C PROV, Janus,...

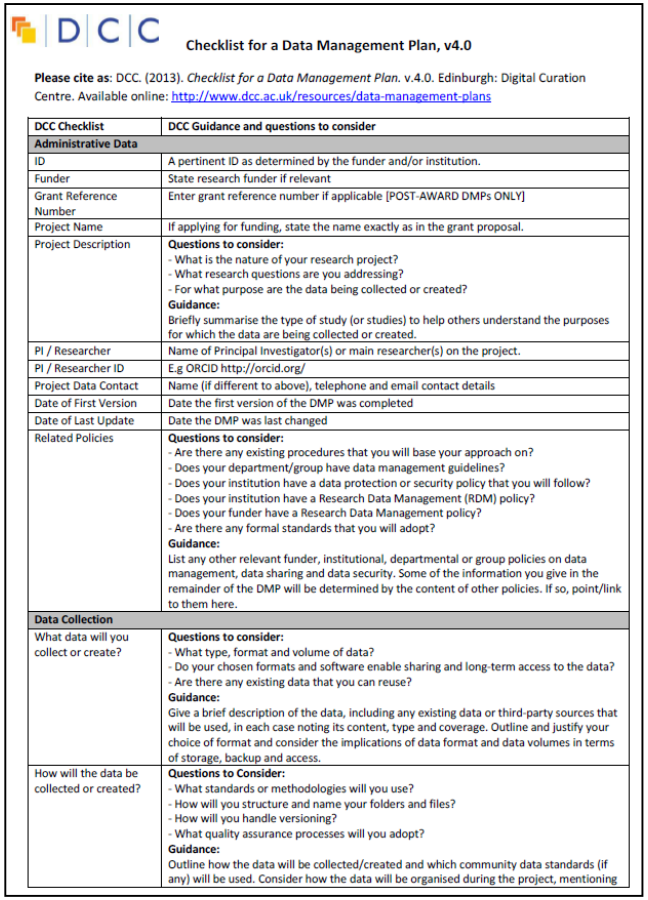
Plato, C3PO,...

FITS, Droid,...

CKAN, Research Compendia, Zenodo...

Research Objects, Context Model, ...

PMF, CDE, noWorkflow,...



DCC

Checklist for a Data Management Plan, v4.0

Please cite as: DCC. (2013). *Checklist for a Data Management Plan.* v.4.0. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/data-management-plans

DCC Checklist	DCC Guidance and questions to consider
Administrative Data	
ID	A pertinent ID as determined by the funder and/or institution.
Funder	State research funder if relevant
Grant Reference Number	Enter grant reference number if applicable [POST-AWARD DMPs ONLY]
Project Name	If applying for funding, state the name exactly as in the grant proposal.
Project Description	**Questions to consider:** - What is the nature of your research project? - What research questions are you addressing? - For what purpose are the data being collected or created? **Guidance:** Briefly summarise the type of study (or studies) to help others understand the purposes for which the data are being collected or created.
PI / Researcher	Name of Principal Investigator(s) or main researcher(s) on the project.
PI / Researcher ID	E.g ORCID http://orcid.org/
Project Data Contact	Name (if different to above), telephone and email contact details
Date of First Version	Date the first version of the DMP was completed
Date of Last Update	Date the DMP was last changed
Related Policies	**Questions to consider:** - Are there any existing procedures that you will base your approach on? - Does your department/group have data management guidelines? - Does your institution have a data protection or security policy that you will follow? - Does your institution have a Research Data Management (RDM) policy? - Does your funder have a Research Data Management policy? - Are there any formal standards that you will adopt? **Guidance:** List any other relevant funder, institutional, departmental or group policies on data management, data sharing and data security. Some of the information you give in the remainder of the DMP will be determined by the content of other policies. If so, point/link to them here.
Data Collection	
What data will you collect or create?	**Questions to consider:** - What type, format and volume of data? - Do your chosen formats and software enable sharing and long-term access to the data? - Are there any existing data that you can reuse? **Guidance:** Give a brief description of the data, including any existing data or third-party sources that will be used, in each case noting its content, type and coverage. Outline and justify your choice of format and consider the implications of data format and data volumes in terms of storage, backup and access.
How will the data be collected or created?	**Questions to Consider:** - What standards or methodologies will you use? - How will you structure and name your folders and files? - How will you handle versioning? - What quality assurance processes will you adopt? **Guidance:** Outline how the data will be collected/created and which community data standards (if any) will be used. Consider how the data will be organised during the project, mentioning

OSF

GitHub

Taverna, VisTrails, Pegasus,...

Jupyter, Zeppelin, ...

Reprozip, Vagrant,...

Docker

Requirements & Limitations

Derived requirements

1. maDMPs must follow a precisely defined schema
   - machine actionability
2. maDMPs must be open
   - incorporate new data types, models and descriptions
3. maDMPs cannot impose limits on technologies
   - experiments implemented using any technology

Derived requirements

4. maDMPs cannot be an evaluation mean per se
5. maDMPs must accommodate needs for manually collected information
6. maDMPs should use closed questions whenever possible and depend on controlled vocabularies

DMP Editor

Data preservation

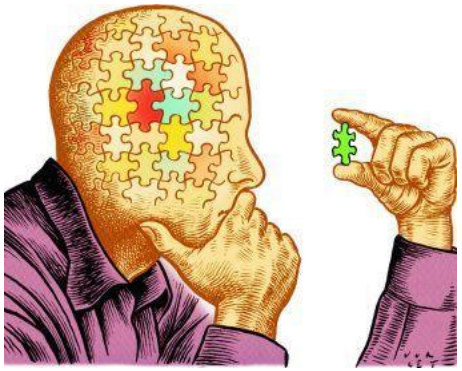
Data Files	Format	Preserved	Duration	Repository
Input data	RAW	☑	10 years	Phaidra
Visualisation	TIF	☑	25 years	Zenodo
Inter. data	CSV	☐	0 years	none

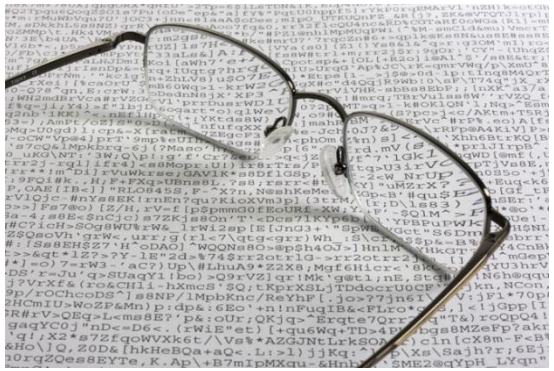
Derived requirements

7. maDMPs must be scalable
   - describe local and distributed experiments
8. maDMPs must link to unique and identifiable entities
   - people
   - repositories
   - licenses

Automation and limitations

- Automatic information
  - collection
  - integration and reasoning
  - validation



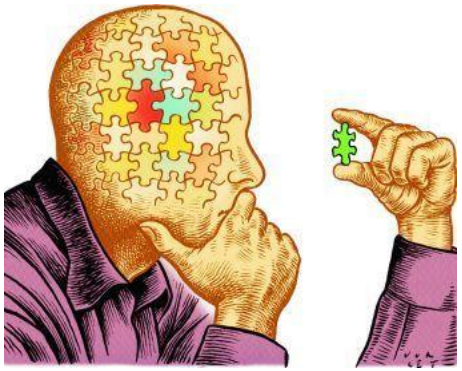


Automation and limitations

- Automatic information collection
  - data characteristics
    - data collection, metadata
  - actions and conditions
    - backups, versioning, licenses

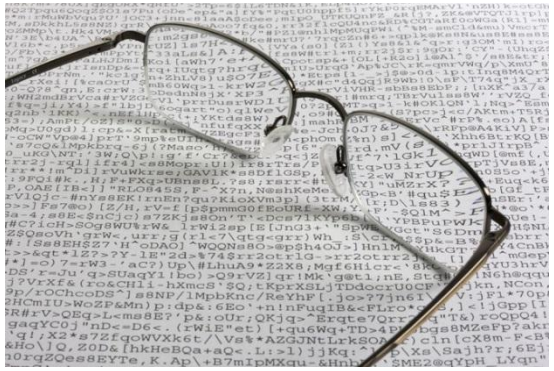
Automation and limitations

- Automatic information integration and reasoning
  - reasoning
  - aggregations and statistics
  - controlled vocabularies for manual inputs



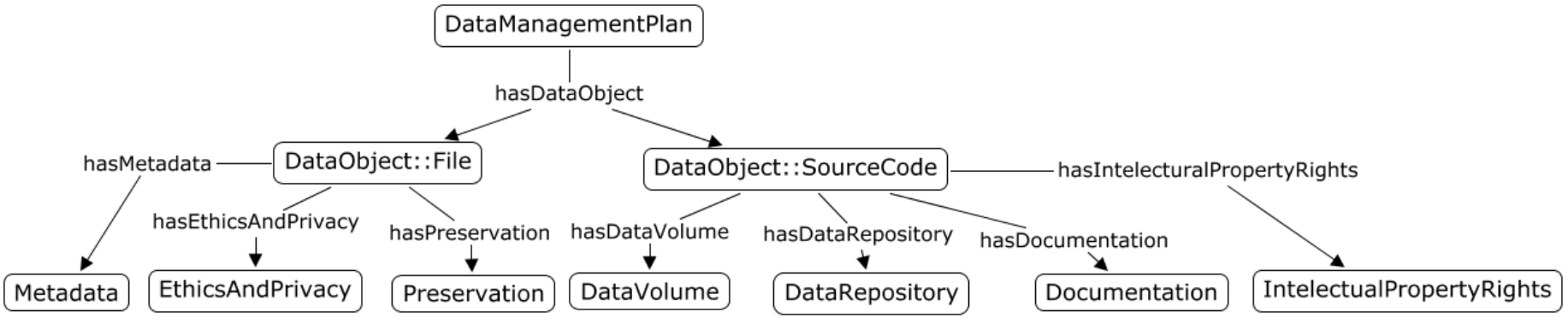
Automation and limitations

- Automatic information validation
  - completeness
  - correctness
  - reality
    - data characteristics
    - actions and conditions

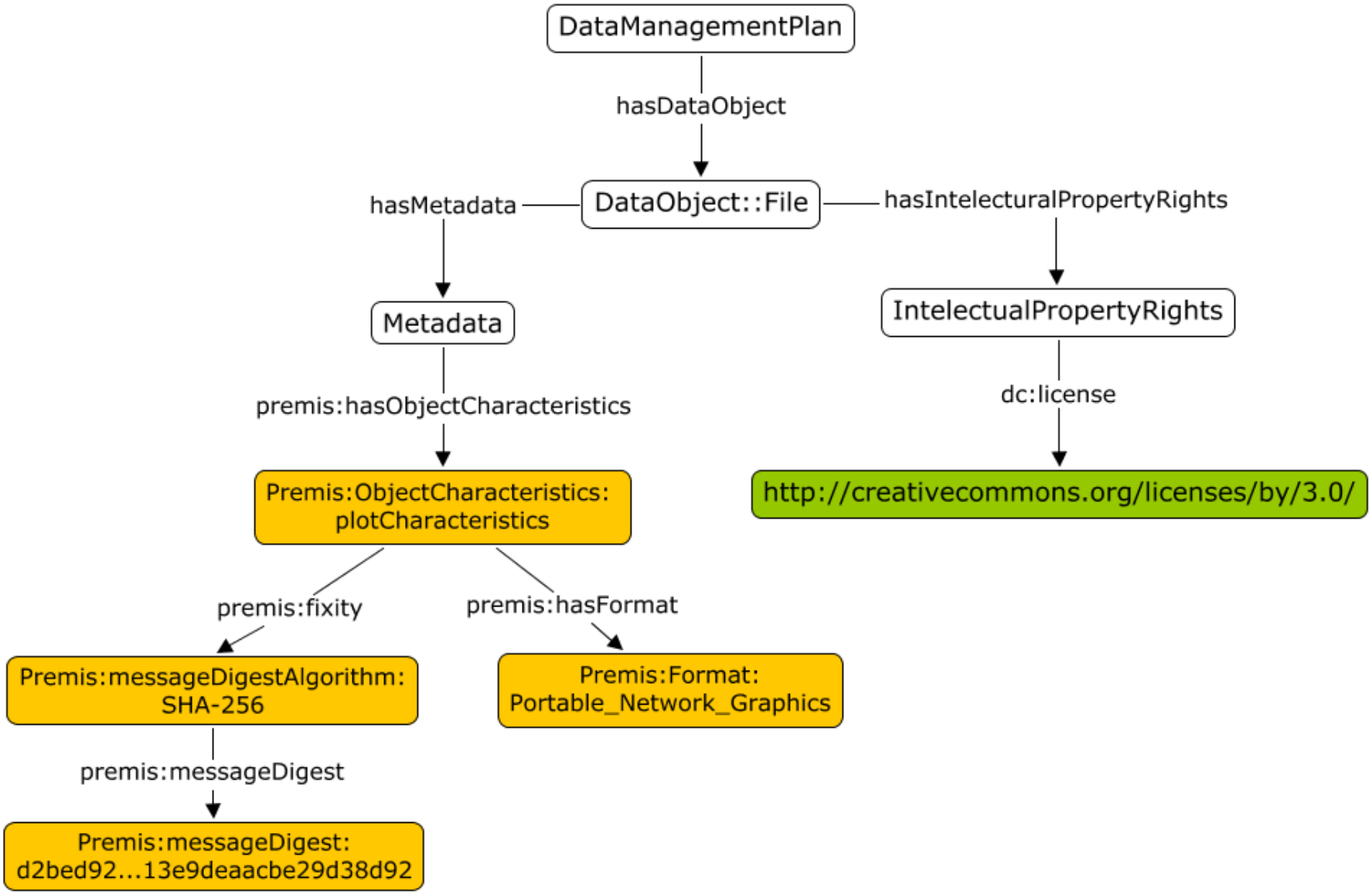


Data model

- Top level vocabulary
- Based on DMP themes
- Extended by domain specific standards
- OWL ontology: https://purl.org/madpms



Example


DataManagementPlan
hasDataObject
DataObject::File
hasMetadata
hasIntelecturalPropertyRights
Metadata
IntelectualPropertyRights
premis:hasObjectCharacteristics
dc:license
Premis:ObjectCharacteristics:
plotCharacteristics
http://creativecommons.org/licenses/by/3.0/
premis:fixity
premis:hasFormat
Premis:messageDigestAlgorithm:
SHA-256
Premis:Format:
Portable_Network_Graphics
premis:messageDigest
Premis:messageDigest:
d2bed92...13e9deaacbe29d38d92

Conclusion and future work

- Mapping of tools to DCC checklist
- Requirements and limitations
- Data model
- Future work
  - iteratively integrate
  - use cases that are gradually more complex
  - Research Data Alliance

tmiksa@sba-research.org