Repository systems

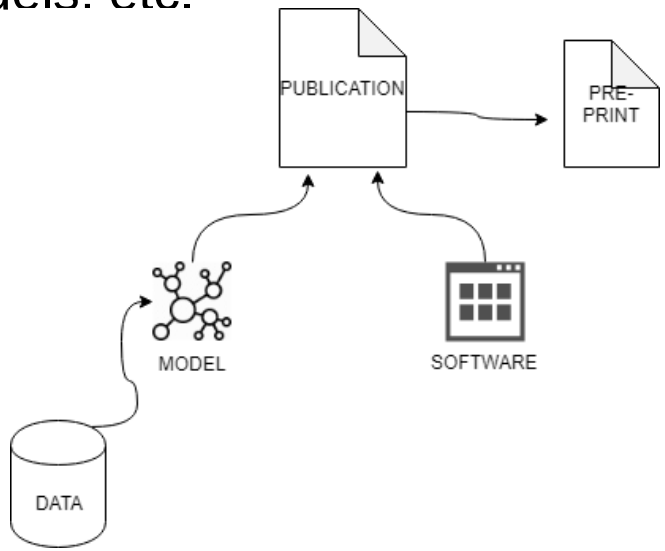
Dr. Tomasz Miksa

Agenda

- Why do we need repositories?
- What is a repository system?
- How to compare repository systems?
- What systems are out there?
- How to make repository contents visible?

Introduction - example

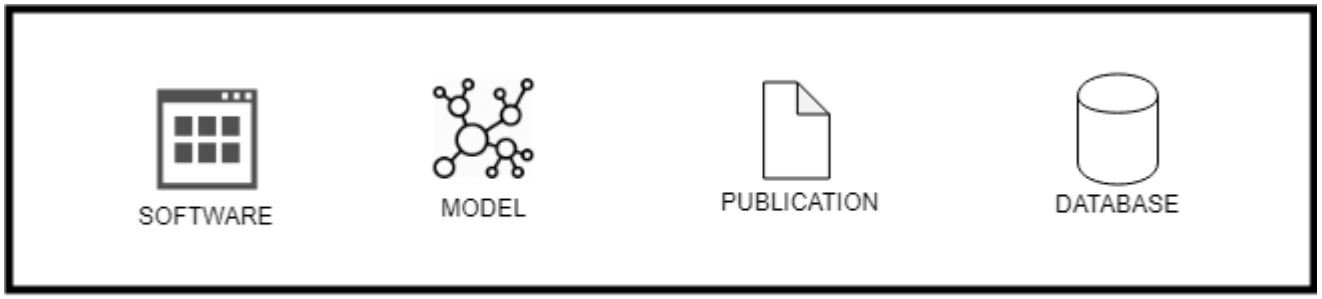
- Science nowadays
  - Continuous or ad-hoc data collection
  - Subsets of data used to train models
  - Software to process data, train models. etc.
  - Research outputs
    - Publication
    - Code
    - Model
    - Data



- ‘If you don’t produce data, you’re not doing science’

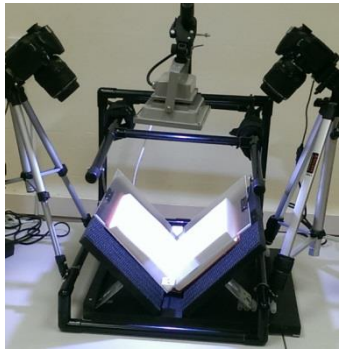
Requirements

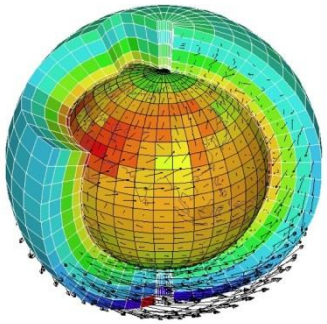
- Repository
  - for all kinds of research material
    - input, output, interim
    - open and closed / sensitive data
  - provides visibility
    - citation, impact
    - FAIR and EOSC compliant
  - Largely transparent to researchers
    - integrates with other infrastructure
  - Trustworthy

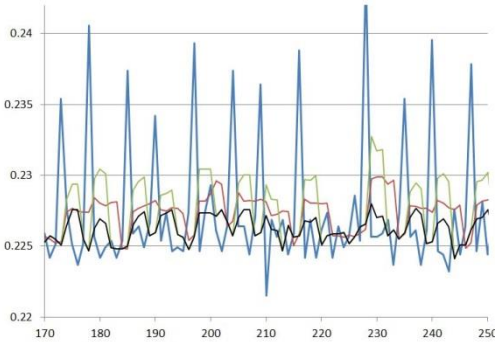


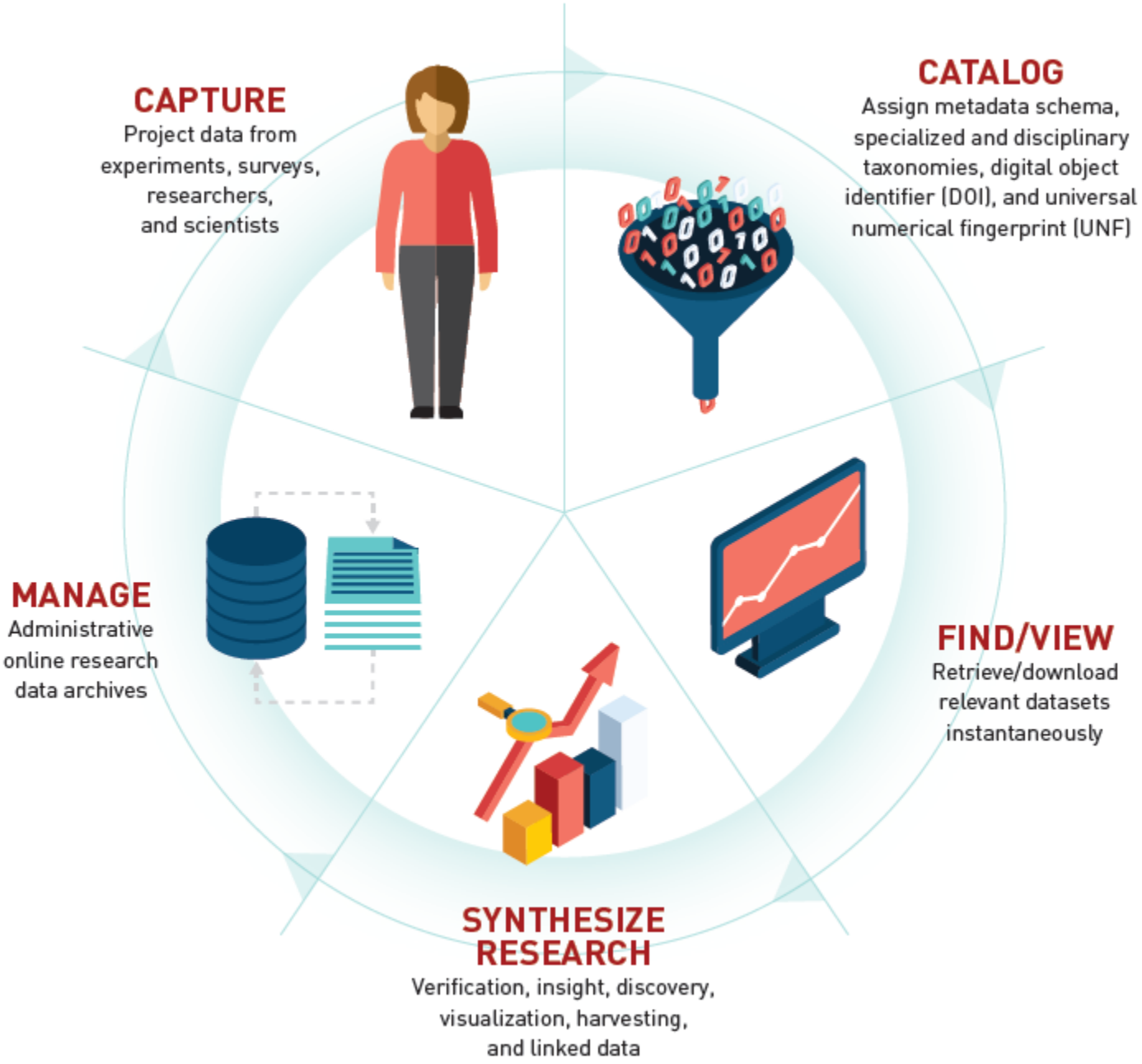
Repositories

- Expectations evolved over time
  - from digitization to preservation of e-Science experiments









http://www.infotoday.com/cilmag/apr16/Uzwyshyn--Research-Data-Repositories.shtml

Phaidra

Search...

Featured collections

E-Books on Demand

In dieser Collection finden Sie die im Rahmen des Services eBooks on Demand (EOD) digitalisierten Bücher der Universitätsbibliothek Wien.

[More] [Collection]

u:scholar

In der u:scholar-Collection von Phaidra finden Sie weltweit frei zugängliche Sie wissenschaftliche Publikationen von Forschenden der Universität Wien.

[More] [Collection]

Digitales Forschungsarchiv Byzanz

Das Ziel des Digitalen Forschungsarchiv Byzanz ist es, das Byzantinische Reich fotografisch möglichst umfangreich zu erfassen und zugänglich zu machen.

[More] [Collection]

UB-Maps

Der Kartenbestand der Fachbereichsbibliothek Geographie und Regionalforschung (Universität Wien) geht historisch bis in die Gründungszeit der Lehrkanzel für Geographie an der Universität Wien im Jahr 1851 zurück.

[Collection]

Digitalisierte Bestände der Österreichischen Zentralbibliothek für Physik

Diese Collection beinhaltet die an der Österreichischen Zentralbibliothek für Physik vorhandenen multimedialen Bestände (Videos, Tonaufzeichnungen etc.) sowie digitalisierte Nachlässe und Sondersammlungen.

[Collection]

650 Jahre Universität Wien

Die Universität Wien feierte 2015 ihr 650. Gründungsjubiläum mit zahlreichen Veranstaltungen.

[More] [Collection]

Contact

phaidra@univie.ac.at

If you have technical problems please contact: support.phaidra@univie.ac.at

« Back to search results

Browse

Title (deu)

Digitalisierte Karten der Fachbereichsbibliothek Geographie und Regionalforschung

Author

Ub-Maps, Projekt

Description (deu)

Beschreibung: Der Kartenbestand der Fachbereichsbibliothek Geographie und Regionalforschung (Universität Wien) geht historisch bis in die Gründungszeit der Lehrkanzel für Geographie an der Universität Wien im Jahr 1851 zurück. Im Jahre 1890 verfügte die Institutsbibliothek bereits über rund 5.000 Kartenblätter, 80 Wandkarten und 115 Atlanten. Die Sammlung enthält Karten aus allen Teilen der Erde, wobei der Schwerpunkt auf dem europäischen Raum liegt (insbesondere Österreich bzw. Länder der Donaumonarchie). Neben historisch wertvollen Altkarten verfügt die Kartensammlung über (aktuelle) Karten aller Art (Landesaufnahmen, thematische Karten, Stadtpläne, Wanderkarten etc.). In den nächsten Jahren werden die historisch bedeutsamen Altkarten sowie urheberrechtsfreie Karten des 20. Jahrhunderts digitalisiert.

Object languages

German

Collection members (185)

15


o:440180 Insula Et Regnum Sardiniae: Longitudine 45. Latitudine 26. milliaria German. complectens ; Cum Gratia et Privil. S.R.I. Vicariat. in partib. Rheni, Sueviae et Iuris Franconici


o:440179 Danmark


o:440178 A geological map of Cyprus

Identifiers

http://phaidra.univie.ac.at/o:423816
Handle: 11353/10.423816

Owner
Projekt UB-Maps

Object Type
COLLECTION

Version
Version 1

Detail page views
443

Object Links
Dublin Core
University of Vienna Metadata

Show full metadata

« Back to search results

Open in new window

Title

Italie: carte générale statistique postale et administrative

Author

Levasseur, Victor

Identifiers

http://phaidra.univie.ac.at/o:440174
AC10837308
Handle: 11353/10.440174

Owner

Projekt UB-Maps

Object Type

PICTURE

Version

Version 1

Detail page views

5

Object Links

View in browser
Download 72 dpi
Download 300 dpi
Download (728.41 MB)
Dublin Core
MODS
EXIF Viewer

Georeference

Latitude, longitude:
41.1833333333333,12.7

Author

Levasseur, Victor

Description

Maßstab in graph. Form (mesures françaises, milles d'Italie). - Statistische Taf. unten links

Description

Bestand der Kartensammlung der Fachbereichsbibliothek Geographie und Regionalforschung, Universität Wien

Description

1:2000000

Description

1 Kt.

Description

53 x 73 cm

Object languages

French

Date

1850

Publisher

Dussillon

Rights

Classification

Italien;Altkarte

bkl: Europa

universität wien ▸ Phaidra

EXIF Viewer for the Object o:440174

IPTC

Coded Character Set	UTF8
Application Record Version	0
By-line	Zeutschel Omniscan 11

EXIF

Subfile Type	Full-resolution Image
Image Width	13080
Image Height	19464
Bits Per Sample	8 8 8
Compression	Uncompressed
Photometric Interpretation	RGB
Strip Offsets	27316
Orientation	Horizontal (normal)
Samples Per Pixel	3
Rows Per Strip	19464
Strip Byte Counts	763767360
X Resolution	600
Y Resolution	600
Planar Configuration	Chunky
Resolution Unit	inches
Software	Adobe Photoshop CS6 (Macintosh)
Modify Date	2015:08:25 14:58:45
Artist	Zeutschel Omniscan 11
Color Space	sRGB
Exif Image Width	13080
Exif Image Height	19464
Padding	(Binary data 2.01 kilobytes)

EMBL-EBI | Services | Research | Training | About us | EMBL-EBI Hinxton

ArrayExpress

Search

Examples: E-MEXP-31, cancer, p53, Geuvadis | advanced search

Home | Browse | Submit | Help | About ArrayExpress | Contact Us | Login

ARRAYEXPRESS / BROWSE / E-MTAB-5993

E-MTAB-5993 - Acetyl-H4 chIP-seq of control, NaBu-treated, and random behavioral clusters of zebrafish 8 dpf larvae

Status	*Submitted on 12 February 2017, last updated on 24 August 2017, released today*
Organism	Danio rerio
Samples (11)	Click for detailed sample information and links to data
Protocols (7)	Click for detailed protocol information
Description	We wanted to analyze the differences in acetyl-H4 content related to inter-individual behavioral variability. We retrieved three groups of samples: The first one was composed by 4 different clusters of zebrafish composed each one by 5 zebrafish larvae. Within each cluster, the behavioral differences across the larvae were minimal, while between the clusters, there were high differences in the behavior. The second one was composed by 4 different clusters of sodium butyrate (NaBu)-treated zebrafish composed each one by 5 larvae. The selection choice was the same as in the first group, but due to their behavioral effect of NaBu, the differences between the clusters were significantly reduced compared to control. The third one was composed by 3 different clusters of 5 zebrafish each randomly selected from the behavioral space. This group is a control of variability not associated to behavior.
Experiment types	ChIP-seq, compound treatment design, innate behavior design
Contact	Angel Carlos Roman <angel.roman@neuro.fchampalimaud.org>
MINSEQE	Exp. design: * Protocols: - Variables: * Processed: * Seq. reads: *
Files	Investigation description: E-MTAB-5993.idf.txt Sample and data relationship: E-MTAB-5993.sdrf.txt Processed data (2): E-MTAB-5993.processed.1.zip, E-MTAB-5993.processed.2.zip Click to browse all available files
Links	ENA - ERP025138, ArrayExpress - E-MTAB-5992 Send E-MTAB-5993 data to GENOMESPACE

http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5993/

Log in | Register | Contact

Climate Change Centre AUSTRIA

Data Centre

Datasets | Organizations | Groups | About

/ Organizations / Wegener Center / ÖKS15 Bias Corrected ... /
rsds_CNRM-CERFACS-CNRM-CM5_RCP...

Create Subset | Go to resource

rsds_CNRM-CERFACS-CNRM-CM5_RCP4.5_r1i1p1_CLMcom-CC...

URL: https://data.ccca.ac.at/dataset/a513367c-a348-4d29-be7f-53ea5089599d/resource/88d350e9-5e91-4922-8d8c-8857553d5d2f/downloa...

Global radiation

Bias corrected (scaled distribution mapping) data of the EURO-CORDEX model CNRM-CERFACS-CNRM-CM5_rcp45_r1i1p1_CLMcom-CCLM4-8-17 using observational data from Global radiation dataset (ZAMG).

Historical and future projection under the RCP4.5 scenario.

Reference period: 1981-2005

Variable

Global radiation

Resource Versions

Resources

- rsds_CNRM-CERFACS-CNRM
- rsds_CNRM-CERFACS-CNRM-CM...
- rsds_CNRM-CERFACS-CNRM-CM...
- rsds_CNRM-CERFACS-CNRM-CM...
- rsds_CNRM-CERFACS-CNRM-CM...
- rsds_CNRM-CERFACS-CNRM-CM...
- rsds_ICHEC-EC-EARTH_RCP4....
- rsds_ICHEC-EC-EARTH_RCP4....
- rsds_ICHEC-EC-EARTH_RCP4....
- rsds_ICHEC-EC-EARTH_RCP4....
- rsds_ICHEC-EC-EARTH_RCP8....
- rsds_ICHEC-EC-

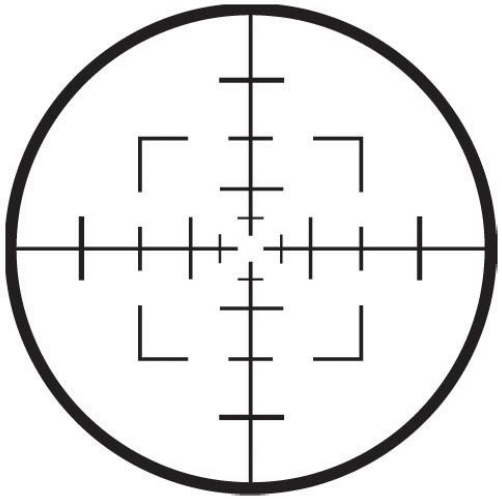
Resource Information

Field	Value
Last updated	September 28, 2016
Created	August 25, 2016
Format	NetCDF
Size	12.2 GiB
Hash MD5	997e2d9fe18e307894e3cc61d362946d
License	Creative Commons Attribution Share-Alike
Format Version	
Resource URI	https://hdl.handle.net/20.500.11756/88d350e9
Anonymous Download	
Newer Version	
Subset of	
Resource (Model) Parameter	
Models	cnrm-cerfacs-cnrm-cm5
Experiment	rcp4.5
Frequency	day
Variables	rsds
Ensemble	r1i1p1

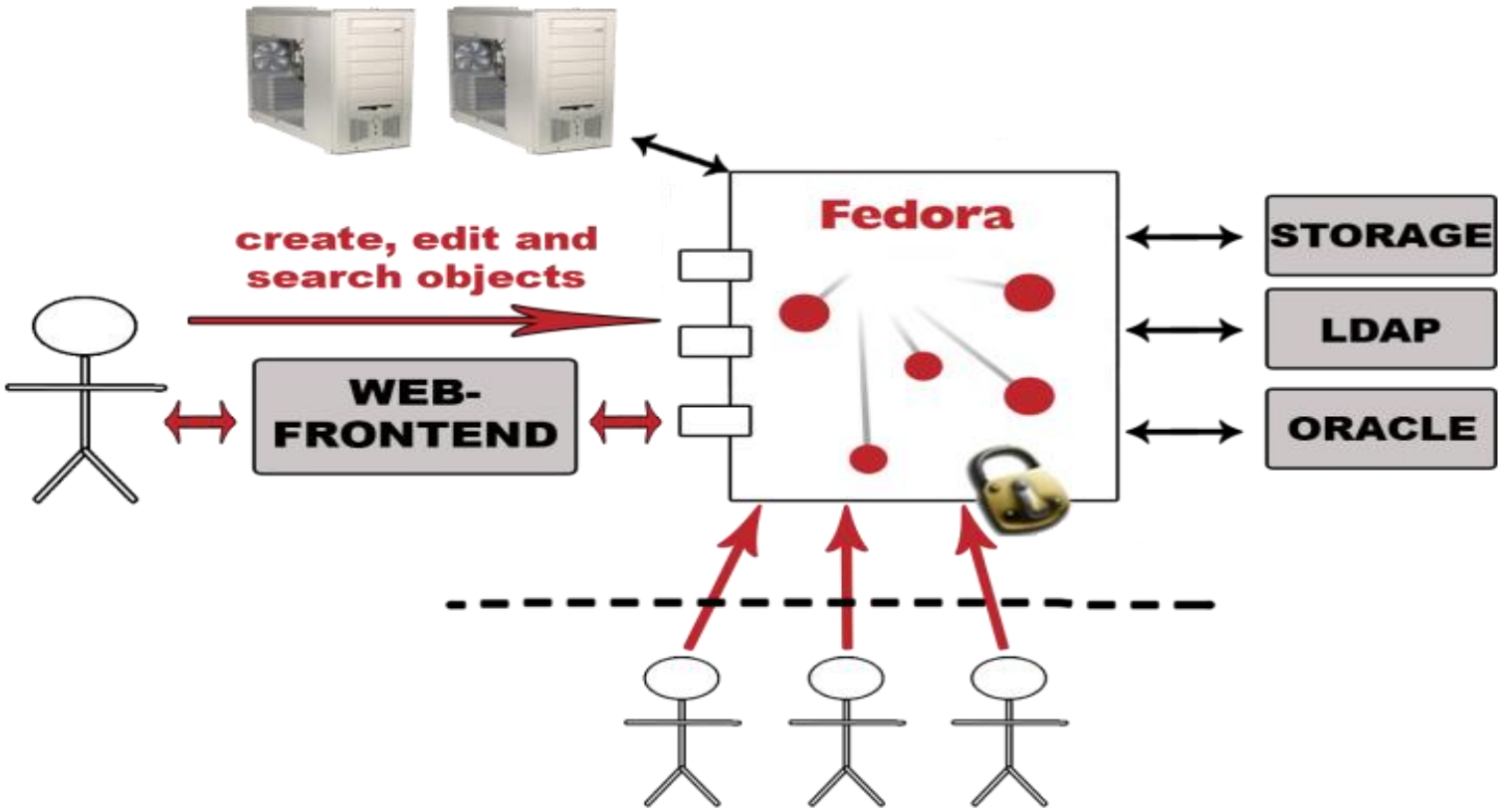
https://data.ccca.ac.at/dataset/oks15-bias-corrected-euro-cordex-models-global-radiation/resource/88d350e9-5e91-4922-8d8c-8857553d5d2f?view_id=eeeb1e17-c707-46eb-bf24-dd5ed169f1c6

Repositories scope

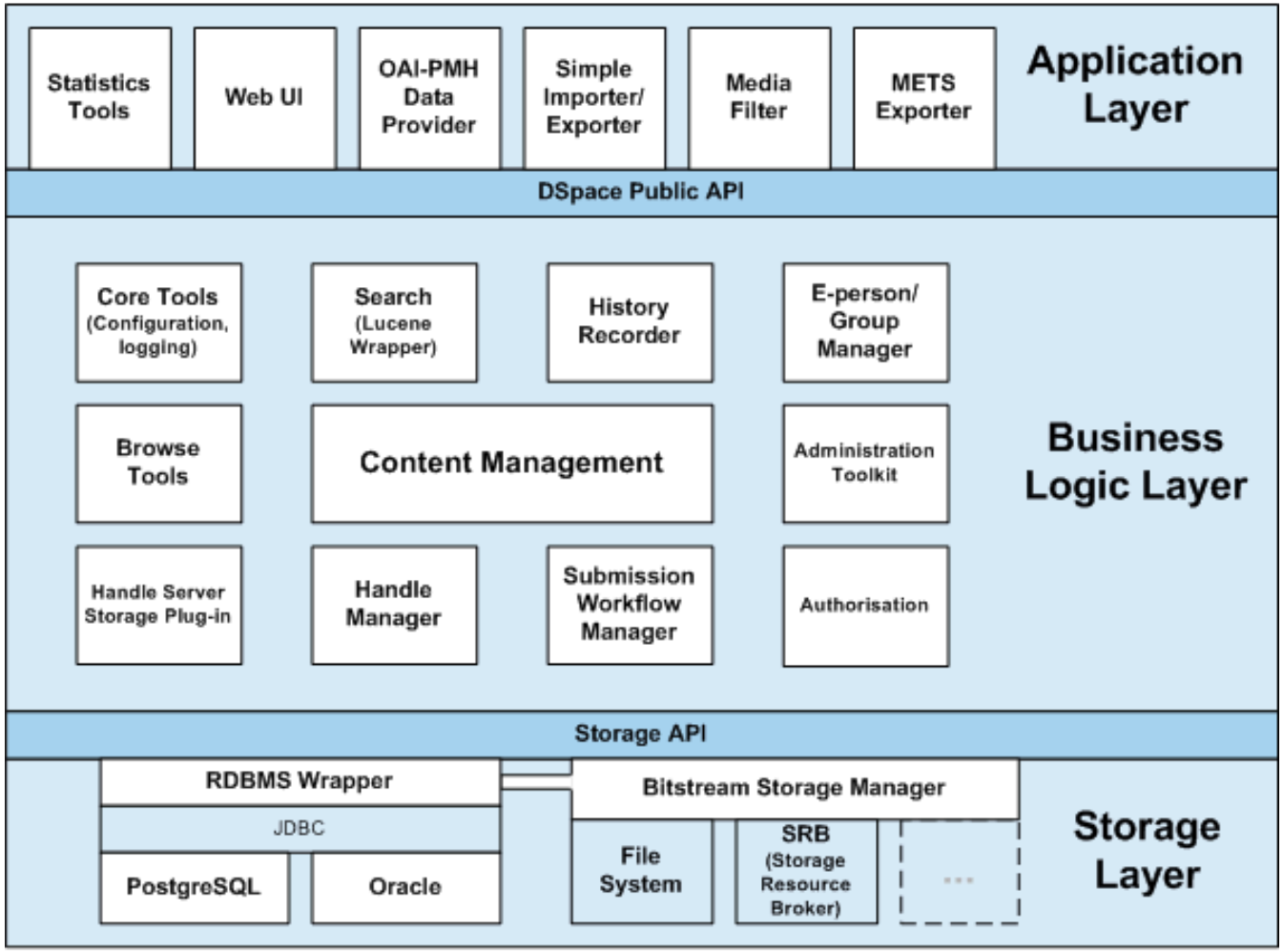
- Specialised
  - disciplinary data
    - e.g. DNA sequencing
- General
  - covering large knowledge areas
    - e.g. social sciences
- Aggregate experts' data
  - globally
  - locally
    - university
    - country



Architecture

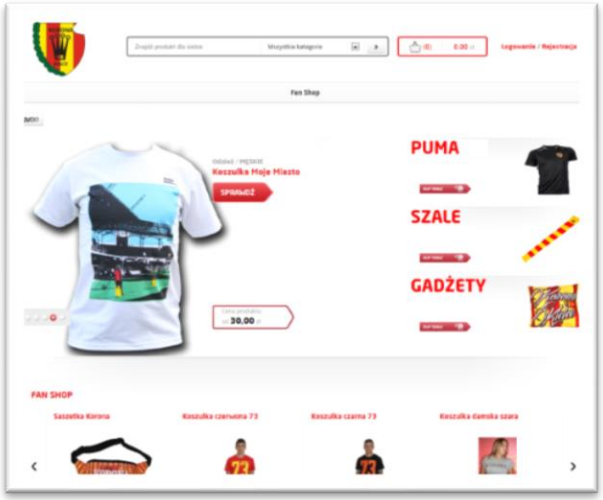


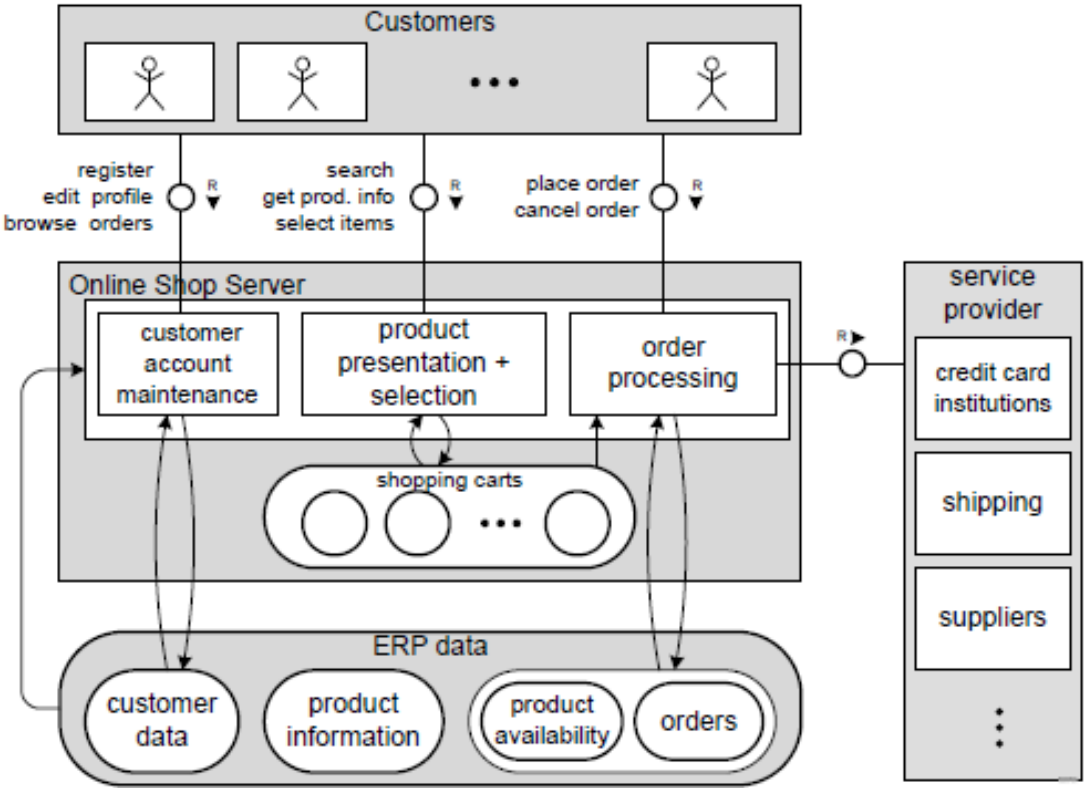
Architecture



Architecture

- Conceptually like any web systems consisting of
  - frontend
  - backend
- Example
  - online shopping





HOW TO COMPARE REPOSITORY SYSTEMS?

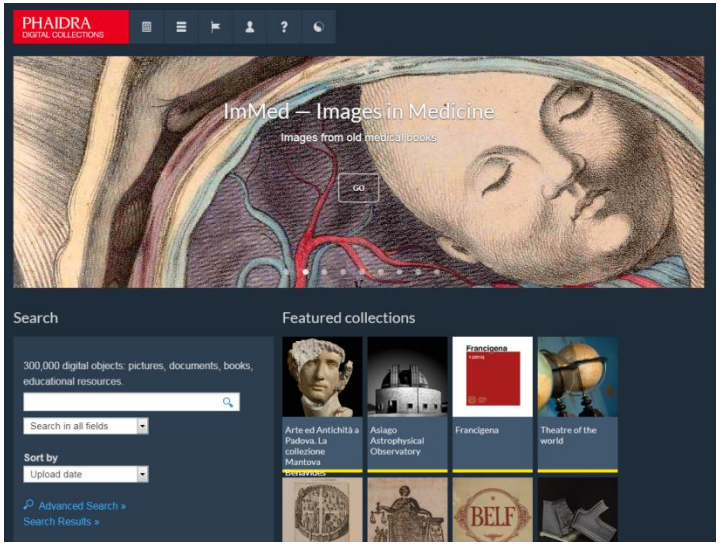
Infrastructure

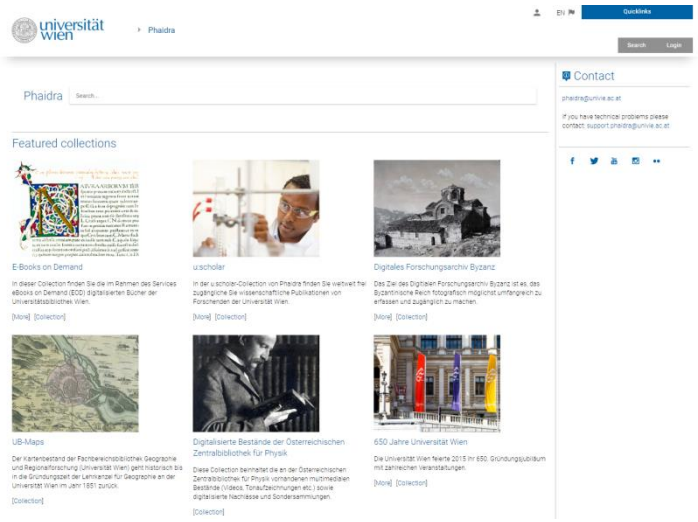
- Locally hosted solution
  - own ICT infrastructure required
  - IT staff required
    - developers, system administrators
- Externally hosted solution
  - outsourcing of infrastructure
  - lack of control where the data is
  - can the external party be trusted?
  - how good is the technical support?
- Open source or Proprietary
  - who owns the code?
  - is it allowed to introduce changes?
- Community support
  - How big? Who is behind it?
  - number of similar instances
  - forum and mailing lists
  - professional support

- Repositories are as trustworthy as institutions behind them

Front-end Design

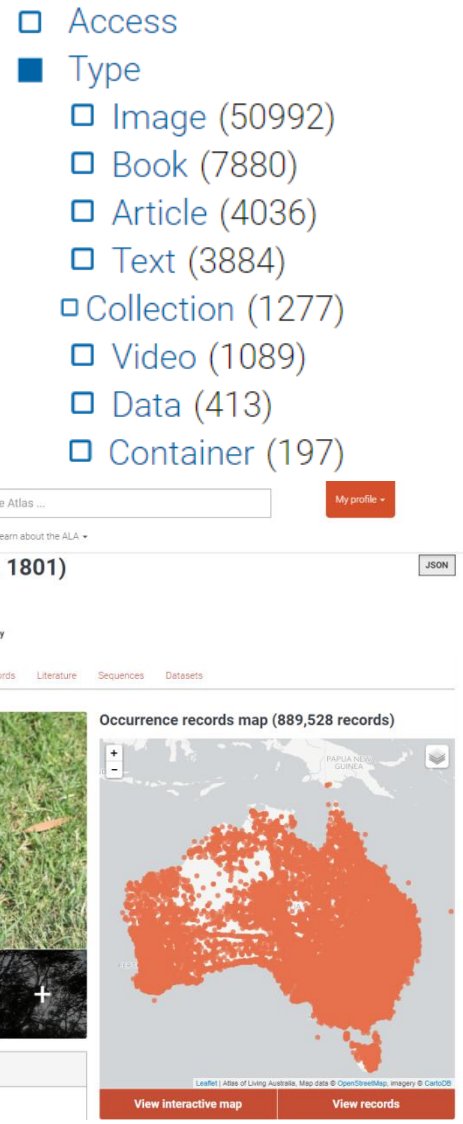
- Out of the box
  or development needed?
  - Fedora Commons is just a backend
- Customisable?
  - branding
- Multi-lingual support?
- Mobile-optimized design?





Content Organization

- Aggregations and Collections
  - proceedings, department outputs, etc.
  - help in navigating the repository
    - faceted search
- Content discovery
- Metadata
  - what are the default standards?
  - how easy to add another standard or new fields?

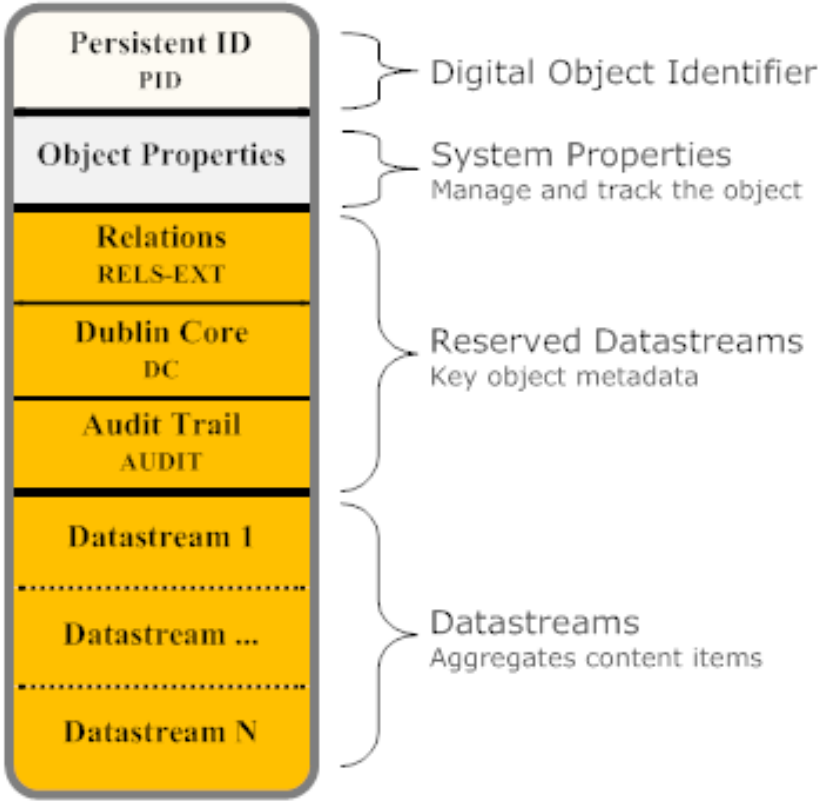


Metadata – Dublin Core

- Originally 15 elements
  http://purl.org/dc/elements/1.1
  - Title
  - Creator
  - Subject
  - Description
  - Publisher
  - Contributor
  - Date
  - Type
  - Format
  - Identifier
  - Source
  - Language
  - Relation
  - Coverage
  - Rights
- DCMI Metadata Terms
  - http://purl.org/dc/terms/
- RFC 5013
  - https://tools.ietf.org/html/rfc5013

Content Organization & Multimedia

- Single object may have many representations
  - (think of maDMPs and DCAT: Dataset and Distribution)
- Content presentation
  - PDF Viewer
  - Video streaming
  - Image previews
  - Audio playback
  - Slideshows



Publication tools

- Customisable submit forms
  - specify required information for a submission
    - metadata
    - license
    - etc.
- Publishing workflow
  - roles
    - editor, reviewer
  - notifications
- Batch processing

Access Control & Authentication

- Data can be
  - Open
  - Shared
  - Closed
- Access control
  - IP ranges, user accounts, Access Control Lists
- Authentication
  - Possible integrations to be considered
    - LDAP
    - System accounts
    - Shibboleth
    - ORCID
  - Embargo periods
- Closed data access control
- Challenge: volatile organisational structure

Reporting

- Needed for feedback and building the case
- Download reports
- Active users
- Google analytics integration

Reported period	Year 2010				
First visit	01 Jan 2010 - 00:52				
Last visit	31 Dec 2010 - 22:12				
	Unique visitors	Number of visits	Pages	Hits	Bandwidth
Viewed traffic *	**<= 6,026** Exact value not available in 'Year' view	**7,619** (1.26 visits/visitor)	**15,045** (1.97 Pages/Visit)	**15,045** (1.97 Hits/Visit)	**4.82 MB** (0.64 KB/Visit)
Not viewed traffic *			**134,506**	**142,025**	**43.93 MB**

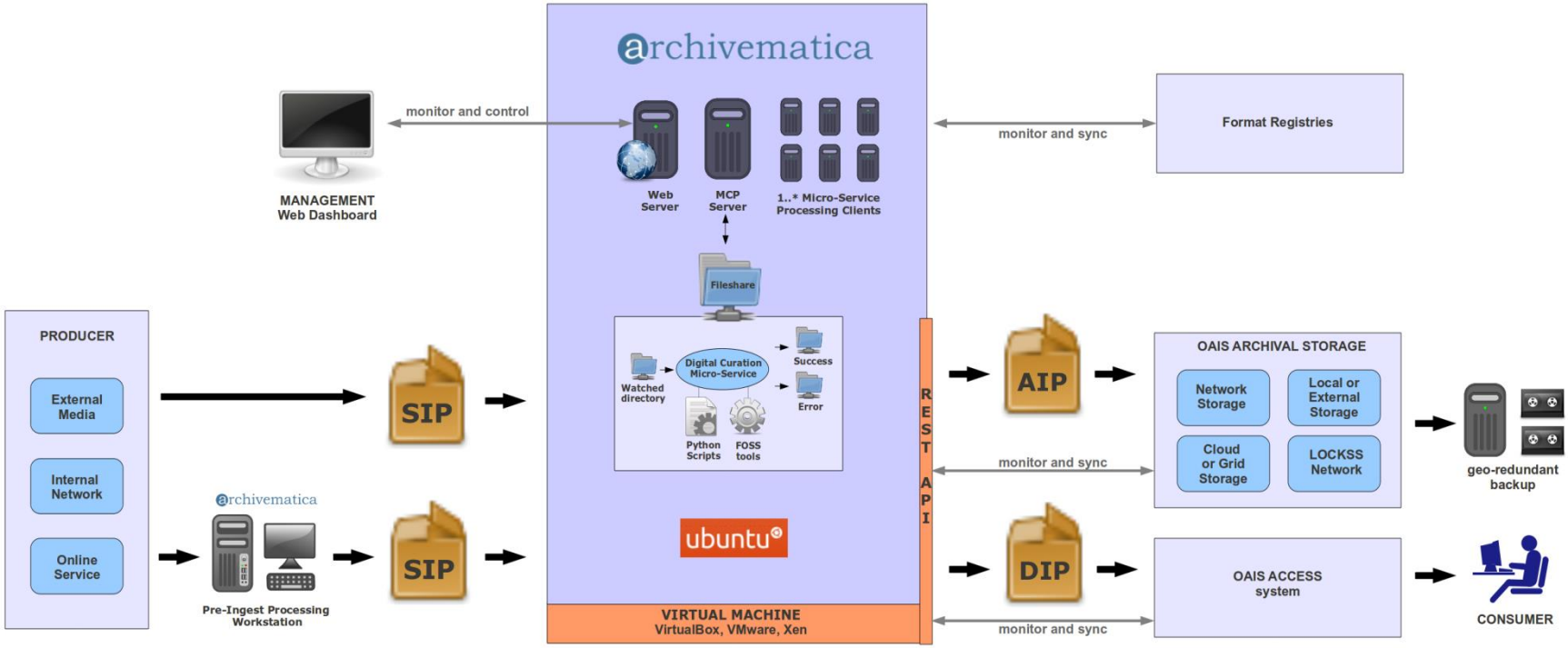
* Not viewed traffic includes traffic generated by robots, worms, or replies with special HTTP status codes.

Preservation

- Back-ups
  - file system backup
  - import / export functionality
- LOCKSS compatibility
  - Lots of Copies Keep Stuff Safe
    - peer to peer network
- Preservation tools
  - format migration tools
  - risk management tools
  - preservation specific metadata collection
    - PREMIS, METS

https://www.loc.gov/standards/premis/
http://www.loc.gov/standards/mets/

Archivematica



archivematica®

Archivematica

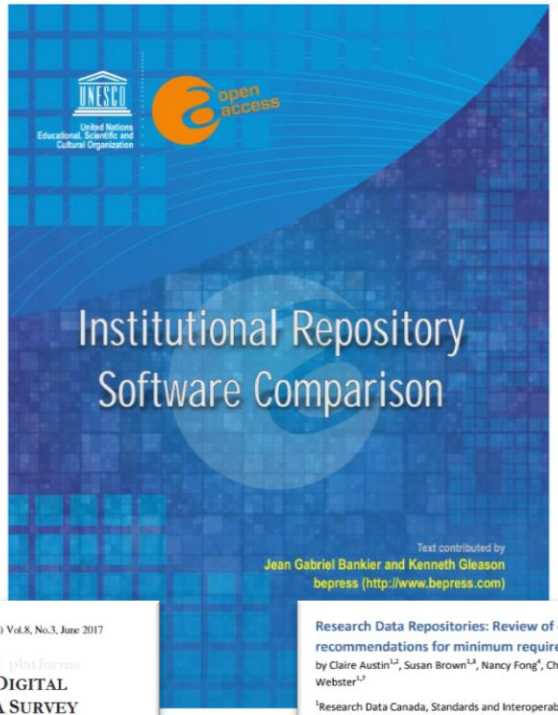
Media type	File formats	Preservation format(s)	Access format(s)	Normalization tool
Audio	AC3, AIFF, MP3, WAV, WMA	WAVE (LPCM)	MP3	FFmpeg
Email	PST	MBOX	MBOX	readpst
Email	Maildir**	Original format	MBOX	md2mb.py
Office Open XML	DOCX, PPTX, XLSX	Original format	Original format	Tool search in progress
Plain text	TXT	Original format	Original format	None
Portable Document Format	PDF	PDF/A	Original format	Ghostscript
Presentation files	PPT	Original format	PDF	Tool search in progress
Raster images	BMP, GIF, JPG, JP2*, PCT, PNG*, PSD, TIFF, TGA	Uncompressed TIFF	JPEG	ImageMagick
Raw camera files/Digital Negative format**	3FR, ARW, CR2, CRW, DCR, DNG, ERF, KDC, MRW, NEF, ORF, PEF, RAF, RAW, X3F	Original format	JPEG	ImageMagick/UFRaw
Spreadsheets	XLS	Original format	Original format	None
Vector images	AI, EPS, SVG	SVG	PDF	Inkscape

https://wiki.archivematica.org/Format_policies

Institutional repository software comparison:
DSpace, EPrints, Digital Commons, Islandora and Hydra

Michel Castagné

University of British Columbia



A comparison of research data management platforms

Architecture, flexible metadata and interoperability

Ricardo Carvalho Amorim, João Aguiar Castro, João Rocha da Silva, Cristina Ribeiro

Abstract Research data management is rapidly becoming a regular concern for researchers, and institutions need to provide them with platforms to support data organization and preparation for publication. Some institutions have adopted institutional repositories as the basis for data deposit, whereas others are experimenting with richer environments for data description, in spite of the diversity of existing workflows. This paper is a synthetic overview of current platforms that can be used for data management purposes. Adopting a pragmatic view on data management, the paper focuses on solutions that can be adopted in the long-tail of science, where investments in tools and manpower are modest. First, a broad set of data management platforms is presented—some designed for institutional repositories and digital libraries—to select a short list of the more promising ones for data management. These platforms are compared considering their architecture, support for metadata, existing programming interfaces, as well as their search mechanisms and community acceptance. In this process, the stakeholders' requirements are also taken into account. The results show that there is still plenty of room for improvement, mainly regarding the specificity of data description in different domains, as well as the potential for integration of the data management platforms with existing research management tools. Nevertheless, depending on the context, some platforms can meet all or part of the stakeholders' requirements.

This paper is an extended version of a previously published comparative study. Please refer to the WCIST 2015 conference proceedings (doi: 10.1007/978-3-319-16486-1)

Ricardo Carvalho Amorim
INESC TEC—Faculdade de Engenharia da Universidade do Porto
E-mail: ricardo.amorim30@gmail.com

João Aguiar Castro
INESC TEC—Faculdade de Engenharia da Universidade do Porto
E-mail: joaoaguiarcastro@gmail.com

1 Introduction

The number of published scholarly papers is steadily increasing, and there is a growing awareness of the importance, diversity and complexity of data generated in research contexts [25]. The management of these assets is currently a concern for both researchers and institutions who have to streamline scholarly communication, while keeping record of research contributions and ensuring the correct licensing of their contents [23, 18]. At the same time, academic institutions have new mandates, requiring data management activities to be carried out during the research projects, as a part of research grant contracts [14, 26]. These activities are invariably supported by software platforms, increasing the demand for such infrastructures.

This paper presents an overview of several prominent research data management platforms that can be put in place by an institution to support part of its research data management workflow. It starts by identifying a set of well known repositories that are currently being used for either publications or data management, discussing their use in several research institutions. Then, focus moves to their fitness to handle research data, namely their domain-specific meta-

International Journal of Computer Science & Engineering Survey (IJCSES) Vol.8, No.3, June 2017

OPEN SOURCE SOFTWARE FOR DIGITAL PRESERVATION REPOSITORIES: A SURVEY

Carlos André Rosa[1], Olga Craveiro[1,2] and Patricio Domingues[1,3]

[1]School of Technology and Management, Polytechnic Institute of Leiria, Portugal
[2]CISUC, Universityof Coimbra, PortugalandAlgoritmi, Universityof Minho, Portugal
[3]Instituto de Telecomunicações, Portugal

ABSTRACT

In the digital age, the amount of data produced is growing exponentially. Governments and institutions can no longer rely on old methods for storing data and passing on the knowledge to future generations. Digital data preservation is a mandatory issue that needs proper strategies and tools. With this awareness, efforts are being made to create and perfect software solutions capable of responding to the challenge of properly preserving digital information. This paper focuses on the state-of-the-art in open-source software solutions for the digital preservation and curation field used to assimilate and disseminate information to designated audiences. Eleven open source projects for digital preservation are surveyed in areas such as supported standards and protocols, strategies for preservation, methodologies for reporting, dynamic of development, targeted operating systems, multilingual support and open source license. Furthermore, five of these open source projects, are further analysed, with focus on features deemed important for the area. Along open source solutions, the paper also briefly surveys the standards and protocols relevant for digital data preservation. The area of digital data preservation repositories has several open source solutions, which can form the base to overcome the challenges to reach mature and reliable digital data preservation.

KEYWORDS

Digital data, preservation, repositories, open source

1. INTRODUCTION

Information preservation can simply be defined as the set of processes to store, index and access information[1]. In recent years, the creation of digital content has grown exponentially. Gantz and Reinsel report that the so called digital universe will grow from 2005 to 2020 by a factor of 300, from 130 exabytes to 40,000 exabytes[2]. They also predict that the whole set of data will double roughly every two years from 2012 to 2020. Digital video is a good example of the current data deluge: the demand for increasing resolutions and higher frame rates, despite all improvements in compression, have substantially increased the size of video files. Smartphones, with all their data sensors, namely photo and video recording capabilities, are also major contributors to the current massive production of data [3]. The Internet of Things (IoT) is poised to generate increasing amount of data, even if IoT middleware can help by reducing the volume of data to store and preserve [4]. The sheer volume of digital information to preserve is immense and will continue to grow over the years. In fact, major trends like Big Data have fostered the perception of digital data as valuable assets, strengthening the need for digital data preservation and henceforth for proper digital repositories [5]. This way, the field of digital information preservation has to address a huge challenge.

DOI:10.5121/ijcses.2017.8302

Research Data Repositories: Review of current features, gap analysis, and recommendations for minimum requirements

by Claire Austin[1,2], Susan Brown[1,3], Nancy Fong[4], Chuck Humphrey[1,5], Amber Leahey[1,6*], Peter Webster[1,7]

[1]Research Data Canada, Standards and Interoperability Committee, [2]Environment Canada, [3]University of Guelph, [4]University of Toronto, [5]University of Alberta, [6]Scholars Portal, [7]Saint Mary's University

*Contact: Amber.Leahey@utoronto.ca

Abstract

Data sharing is increasingly recognized as integral to scientific research and publishing. This requires informed and thoughtful preparation from initial research planning to collection of data/metadata, interoperability, deposit in data repositories, and curation. Research Data Canada (RDC) is a collaborative, non-government organization that promotes access to and preservation of Canadian research data. The RDC Standards and Interoperability Committee (RDC-SINC) surveyed 32 Canadian and International online data platforms for storage, data transfer, curation activities, preservation, access, and sharing features. We developed a checklist to compare criteria and features between platforms. The survey revealed a heterogeneity of features and services across platforms, non-standardized use of terms, uneven compliance with relevant standards, and a paucity of certified data repositories. Recommendations for online digital infrastructure development to meet evolving researcher and end-user needs centre around persistent identification and citation of datasets, data reliability, version control, metadata, data sharing, privacy controls, long-term preservation of data, and certification of data repositories. We identified a need in Canada for investment in an integrated, comprehensive national digital infrastructure for research data.

Keywords

Data sharing, data publication, digital repository, data interoperability, data standards, data deposit, digital infrastructure

INTRODUCTION

Research data sharing is increasingly recognized as an essential component of scholarly and scientific research. Increased sharing improves the ability to reproduce results, replicate findings, and generate new knowledge (Parr & Cummings, 2005; Hernan & Wilcox 2009; Peng 2011; Poisot et al. 2013; Stodden et al. 2014, 2015). Although some disciplines (e.g., astronomy) have a long established practice of sharing and citing scientific data sets (CODATA-ICSTI, 2013), a very large number of researchers are still very reluctant to do so. Perceived risks in data sharing sometimes put forth by researchers, such as damage to the researcher's reputation, misinterpretation of the data, or misappropriation of the data (CODATA-ICSTI, 2013), all immediately disappear the moment the data are properly managed and documented. Some surveys have found that approximately half of

Common survey and report shortcomings

- Surveys not necessarily made by people who will run the systems
- Simplifications
  - ‘customisable’ metadata
    - ‘XML programming’ vs admin interface
- Not up to date
  - systems evolve fast
- Superficial
  - deployment mode - sometimes both options exist
    - Zenodo as a service or open source at GitHub (Invenio)
  - community ‘support’
    - check how many posts there are and how fast people got answers
- Currently no survey to compare them all

Best way is to get your hands dirty

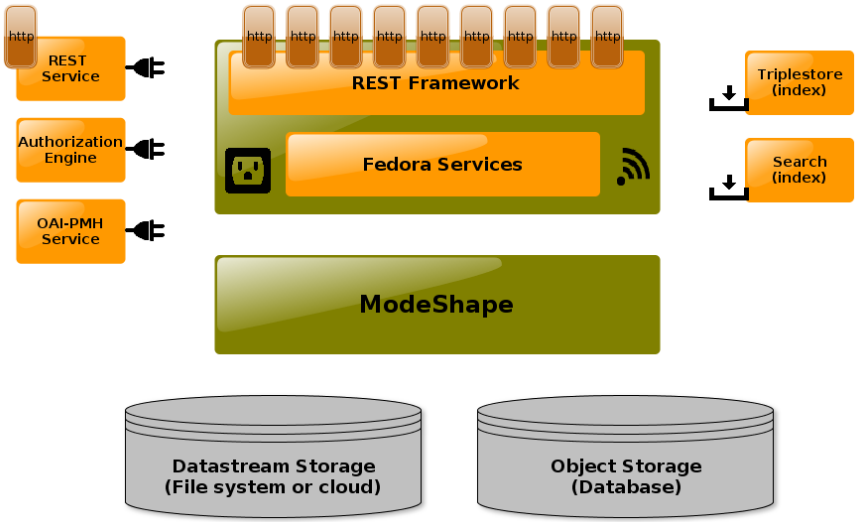
- Sort out the basics
  - What is the purpose of the repository?
  - Do I need more than one repository?
  - Which functionality is a must and which one is nice to have?
  - How should the system integrate with the existing infrastructure?
- Browse through
  - official websites, wiki pages, GitHub issues, mailing lists
- Contact people who already have an instance
- Make test deployment of few systems
  - install
  - configure
  - populate with sample data
  - evaluate

(non-exhaustive overview)

WHAT REPOSITORY SYSTEMS ARE OUT THERE?

Fedora based

- NOT a Linux OS distribution!
- Fedora commons provide backend only
  - Content model
  - RDF linked data
  - Persistent identifier
  - Versioning
- Requires further development



DSpace

- Open source
- Popular
  - scholarly publication repositories
- Large open source community
  - variety of support and consultancy providers
- Aims for ‘turnkey’ local installation, but can be complex to set up and maintain if customisation is required

Dataverse

- Dataverse
  - Driven by Harvard and DANS (NL)
  - own instance -> become part of community
  - Flexible API
- Invenio
  - Powers Zenodo
  - Developed at CERN
  - Integrates with GitHub
  - Recognized by the European Commision

CKAN - Comprehensive Knowledge Archive Network

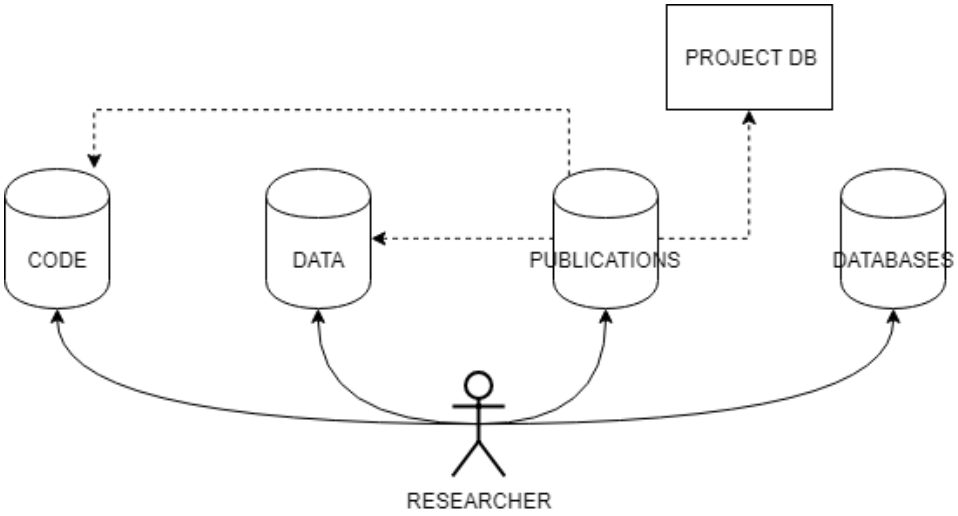
- Repository for data
  - customisable
  - well-established open source community
  - wide take-up in government sector
    - e.g. data.gv.at
- Cons:
  - Limited access restrictions
  - Lacks support for OAI-PMH

Repository software - summary

- Often specialised in one kind of data
  - e.g. GitLab good for software code
- Skewed towards a specific domain (which it originates from)
  - e.g. DSpace used in humanities – focus on documents
  - e.g. CKAN for open (government) data – poor access restrictions mechanisms
- Always a trade-off
  - Some functionality works out of the box, some other must be added
    - Additional package, configuration or development
- All have world-wide community
  - e.g. Dataverse driven by Harvard, Invenio by CERN, CKAN by US administration, etc.

Repository software - summary

- Supporting community usually has a crucial impact on the decision
- Many institutions only tick off a box ‘we have a repository’
- There may be no single system to address all requirements



EXTERNAL VISIBILITY

External visibility

- ‘Having a repository’ is not enough
  - Contents must be discoverable and FAIR
    - Integration with hubs
    - Metadata following standards and machine-actionability

Cargo cult

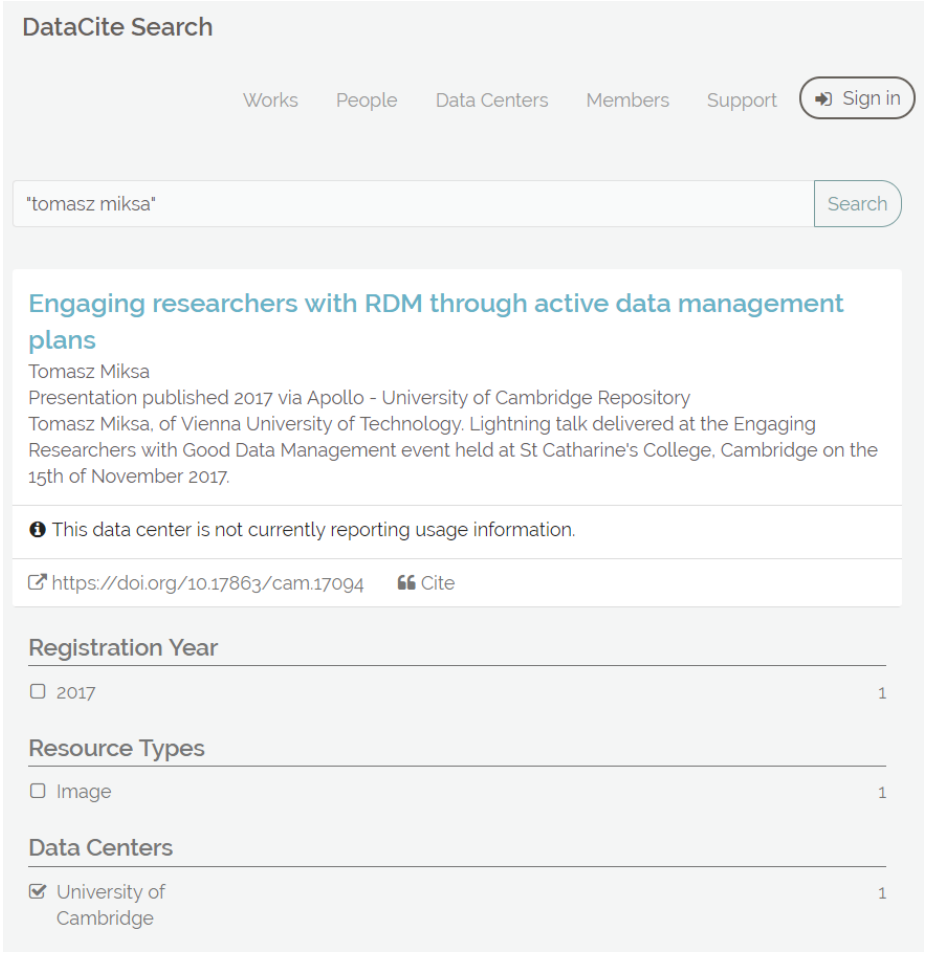
DOI registration bodies

- DataCite
  - DOI registration body
    - Handle based identifier
    - Metadata
    - Provides a range of identifiers to repositories
  - Supported by public sector, e.g. DCC, CERN, ANDS
  - National desk established at the TU Wien

- Crossref
  - DOI registration body
  - Supported by private sector, e.g. Elsevier

DOI registration bodies - example

- Example shows
  - Document located at Cambridge Repository
  - Document has a DOI
  - DOI was minted by Data Cite
  - Data Cite has metadata about each DOI
- Data Cite provides access to metadata registry



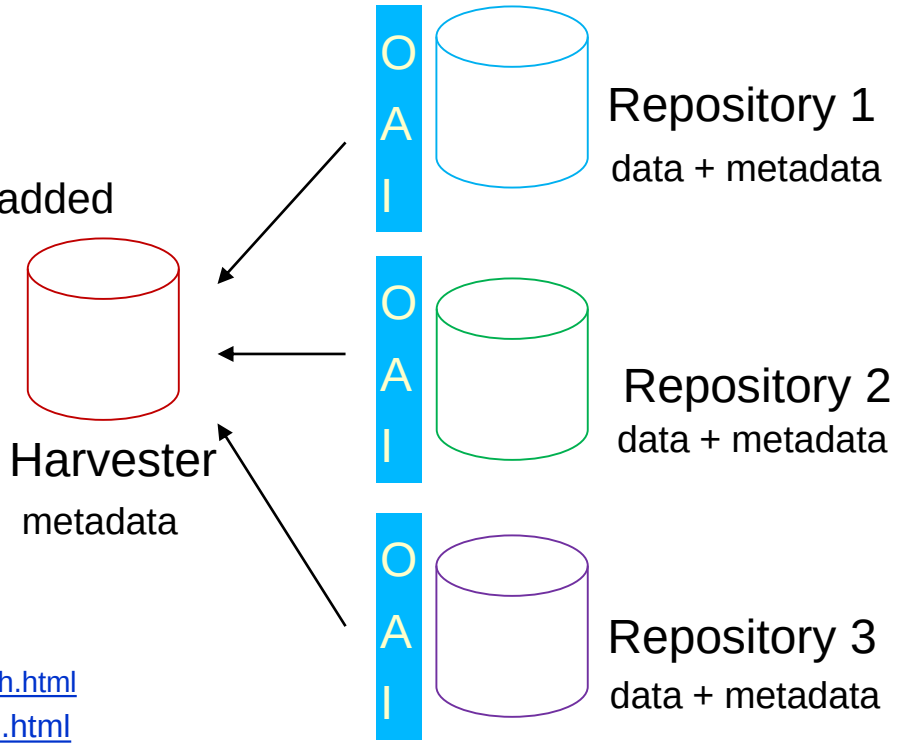
https://search.datacite.org

Interoperability protocols

- OAI-PMH - Open Archives Initiative Protocol for Metadata Harvesting
  - query to discover repository contents
  - only for metadata
  - not for depositing
- SWORD - Simple Web-service Offering Repository Deposit
  - deposit to multiple repositories at once
  - deposit by third party systems (e.g. lab equipment)

OAI-PMH

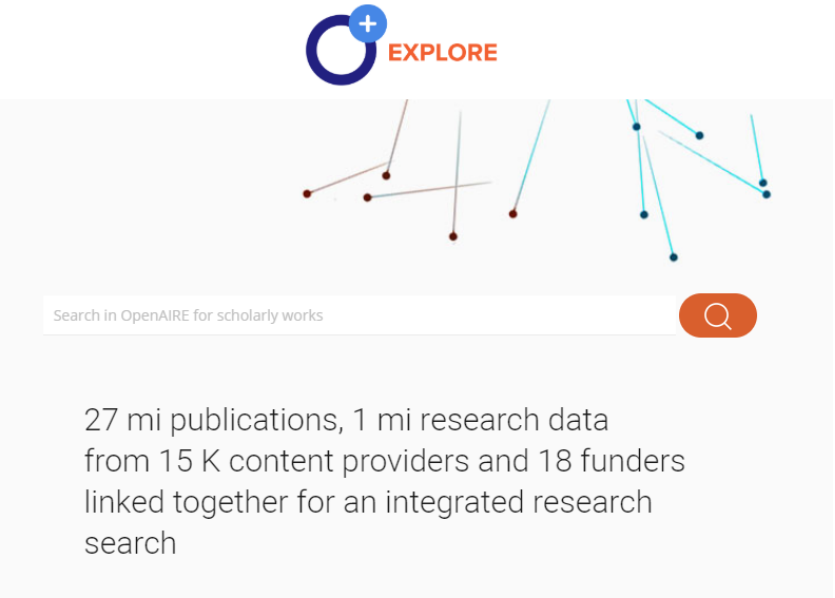
- Data remains within a repository
- Harvester aggregates metadata
- Useful to aggregate data
  - e.g. for domain, country, institution
- Dublin Core by default
  - Other Metadata standards can be added
    - OpenAIRE requires DataCite



https://guidelines.openaire.eu/en/latest/data/use_of_oai_pmh.html
http://www.openarchives.org/OAI/openarchivesprotocol.html

OpenAIRE

- Open Access Infrastructure for Research in Europe
  - Launched 2009 by European Commission
  - Promotes Open Access
- Network of experts
- Technical infrastructure
  - Harvest research output
    - Data, publications
    - Link it
    - Monitor
  - Zenodo



https://explore.openaire.eu

OpenAIRE - compliance

- Repository must have an OAI-PMH endpoint
- Metadata in DataCite format

Table 1: DataCite Mandatory Properties

ID	Property	Obligation
1	Identifier (with mandatory type sub-property)	M
2	Creator (with optional given name, family name, name identifier and affiliation sub-properties)	M
3	Title (with optional type sub-properties)	M
4	Publisher	M
5	PublicationYear	M
10	ResourceType (with mandatory general type description sub-property)	M

Table 2: DataCite Recommended and Optional Properties

ID	Property	Obligation
6	Subject (with scheme sub-property)	R
7	Contributor (with optional given name, family name, name identifier and affiliation sub-properties)	R
8	Date (with type sub-property)	R
9	Language	O
11	AlternateIdentifier (with type sub-property)	O
12	RelatedIdentifier (with type and relation type sub-properties)	R
13	Size	O
14	Format	O
15	Version	O
16	Rights	O
17	Description (with type sub-property)	R
18	GeoLocation (with point, box and polygon sub-properties)	R
19	FundingReference (with name, identifier, and award related sub-properties)	O

https://guidelines.openaire.eu/en/latest/

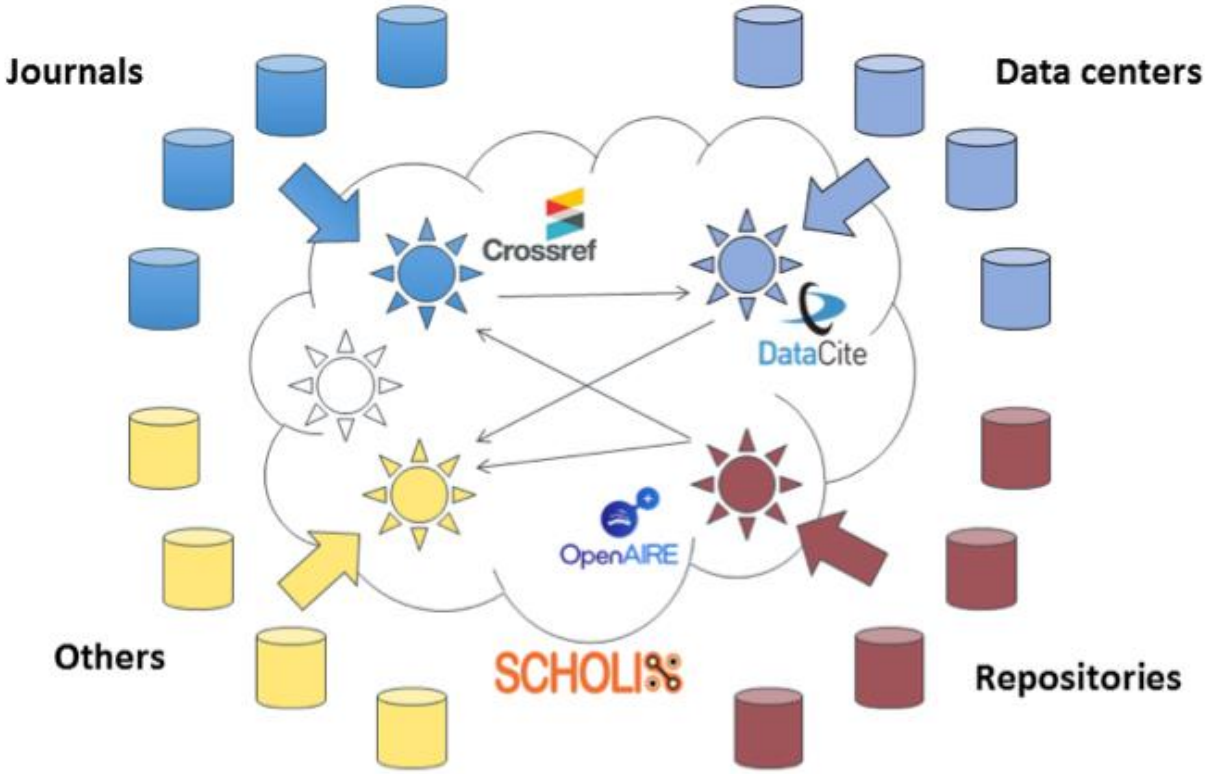
https://schema.datacite.org/meta/kernel-4.2/doc/DataCite-MetadataKernel_v4.2.pdf

OpenAIRE

- Check what kind of sources are indexed and how metadata is collected
  - https://www.openaire.eu/aggregation-and-content-provision-workflows
- What else openAIRE does
  - https://www.openaire.eu/faqs

Scholix

- Goal
  - exchange links between publications and data

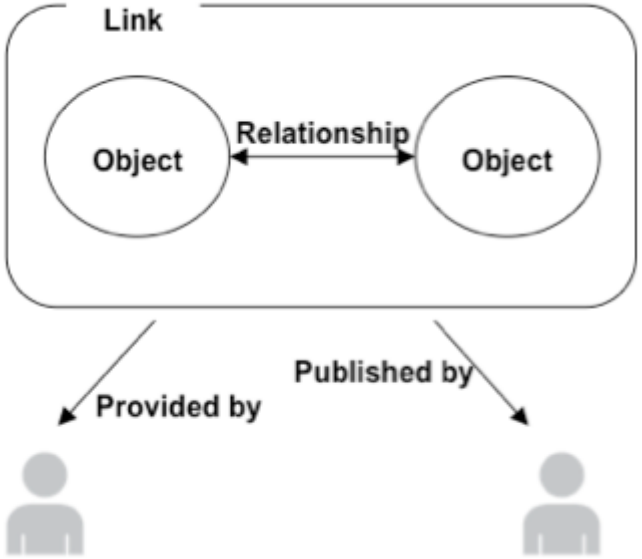


http://www.dlib.org/dlib/january17/burton/01burton.html

Scholix

- At the core of the conceptual model is the *link* between two *objects*.
- Main focus: literature and data.
  - theoretically also: software, algorithms, models, protocols, tweets, comments, and so on.
  - practically: not for the time being the focus
- To become a contributor
  - Feed your data-literature link information to an existing Scholix hub using your existing community standards
    - e.g. OpenAIRE or DataCite registries
- To retrieve links
  - http://api.scholexplorer.openaire.eu/v2/ui/

Scholix



Link Information Package

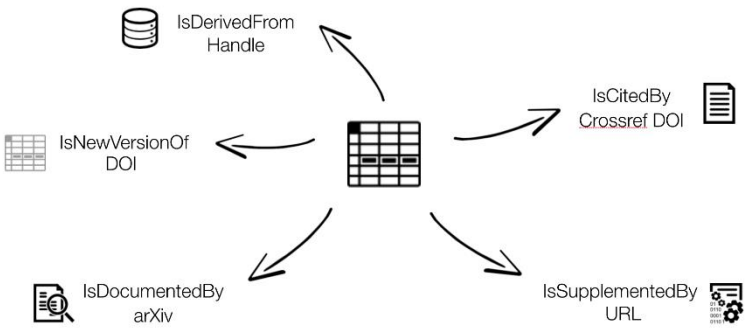
Publication Date (1)
Link publisher (1)
Link Provider (0..1)
Relationship Type (0..1)
Inverse Relationship Type (0..1)

Source Object

PID (1..N)
Object Type (1)
Object Sub-Type (0..1)
Title (0..1)
Author (0..N)
Publication date (0..1)
Publisher (0..1)

Target Object

PID (1..N)
Object Type (1)
Object Sub-Type (0..1)
Title (0..1)
Author (0..N)
Publication Date (0..1)
Publisher (0..1)



http://api.scholexplorer.openaire.eu/v2/ui/

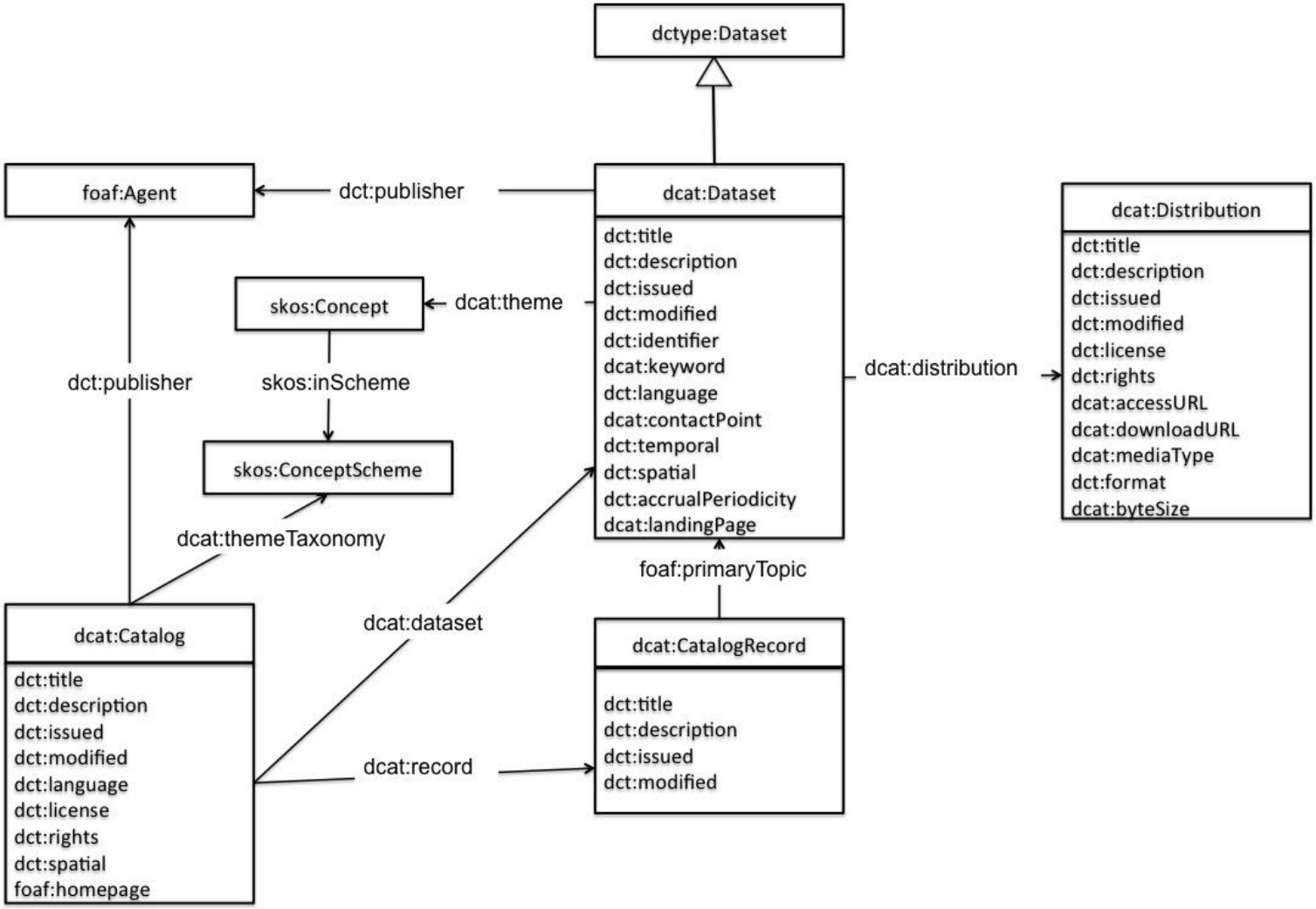
Scholix was developed at RDA

https://www.rd-alliance.org/groups/rdawds-scholarly-link-exchange-scholix-wg

DCAT

- W3C Data Cataolog Vocabulary (DCAT)
  - RDF vocabulary for interoperability between data catalogues
  - decentralized publishing
  - facilitates federated dataset search
- Relaxed constraints
  - Most fields are optional
- No specific deployment method
  - RDF via SPARQL, embedded in HTML, serialised to RDF/XML or Turtle, etc.
- Mostly used in open governmental data repositories

DCAT



Specification and examples: https://www.w3.org/TR/vocab-dcat/

DCAT Application Profile

- **DCAT profile** is a specification that adds additional constraints
- Application Profile for open data in Europe
  - https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe/about

4.3. Dataset

4.3.1. Mandatory properties for Dataset

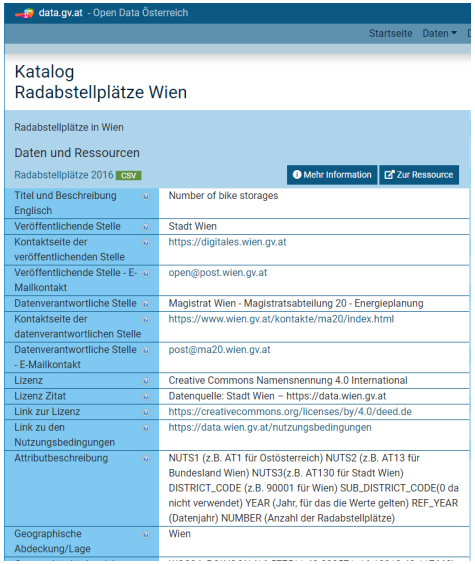
Property	URI	Range	Usage note	Card
description	dct:description	rdfs:Literal	This property contains a free-text account of the Dataset. This property can be repeated for parallel language versions of the description.	1..n
title	dct:title	rdfs:Literal	This property contains a name given to the Dataset. This property can be repeated for parallel language versions of the name.	1..n

4.3.2. Recommended properties for Dataset

Property	URI	Range	Usage note	Card
contact point	dcat:contactPoint	vcard:Kind	This property contains contact information that can be used for sending comments about the Dataset.	0..n
dataset distribution	dcat:distribution	dcat:Distribution	This property links the Dataset to an available Distribution.	0..n
keyword/ tag	dcat:keyword	rdfs:Literal	This property contains a keyword or tag describing the Dataset.	0..n
publisher	dct:publisher	foaf:Agent	This property refers to an entity (organisation) responsible for making the Dataset available.	0..1
theme/ category	dcat:theme, subproperty of dct:subject	skos:Concept	This property refers to a category of the Dataset. A Dataset may be associated with multiple themes.	0..n

DCAT - example

- Add .RDF to links in data.gv.at to get RDF metadata
  - https://www.data.gv.at/katalog/dataset/af8e02b6-1e03-4464-a69b-8533d8703ffe
  - https://www.data.gv.at/katalog/dataset/af8e02b6-1e03-4464-a69b-8533d8703ffe**.rdf**



DCAT - example


```
<dcat:Dataset rdf:about="https://www.data.gv.at/dataset/af8e02b6-1e03-4464-a69b-8533d8703ffe">
    <dct:title>Radabstellplätze Wien</dct:title>
    <dct:description>Radabstellplätze in Wien</dct:description>
    <dcat:keyword>fahrrad</dcat:keyword>
    <dcat:keyword>räder</dcat:keyword>
    <dcat:keyword>verkehr</dcat:keyword>
    <dcat:keyword>fahrräder</dcat:keyword>
    <dcat:distribution>
      <dcat:Distribution rdf:about="https://www.data.gv.at/dataset/af8e02b6-1e03...">
        <dct:title>Radabstellplätze 2016</dct:title>
        <dct:format>CSV</dct:format>
        <dcat:accessURL rdf:resource="https://www.wien.gv.at/gogv/l9radabstellplaetze2016"/>
      </dcat:Distribution>
    </dcat:distribution>
  </dcat:Dataset>
```

SPARQL endpoint - example

EU Open Data Portal

- Retrieve all the resources from a dataset with a title that contains specific words (eg.'Vienna')

SPARQL

You can search for the metadata stored in the EU Open Data Portal triple store by using the SPARQL endpoint query editor below.

Namespaces *

```
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX odp:  <http://data.europa.eu/euodp/ontologies/ec-odp#>
PREFIX dc: <http://purl.org/dc/terms/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
```

SPARQL query *

```
SELECT  ?DatasetTitle  ?Publisher ?ResourceDescription WHERE { graph ?g  {?DatasetURI a dcat:Dataset;
dc:publisher ?Publisher; dc:title ?DatasetTitle; dcat:distribution ?Resource. ?Resource dc:description ?
ResourceDescription. FILTER(regex(?DatasetTitle,"Vienna", "i")) } }      LIMIT 10
```

https://data.europa.eu/euodp/en/linked-data

SPARQL endpoint - example

EU Open Data Portal

DatasetTitle	Publisher	ResourceDescription
"Glossary City of Vienna"	http://publications.europa.eu/resource/authority/corporate-body/CNECT	"Data archive containing files in the following formats: application/xml"
"University of Vienna Termbanks"	http://publications.europa.eu/resource/authority/corporate-body/CNECT	"Data archive containing files in the following formats: application/xml"
"Audioguide for the Military History Museum in Vienna"	http://publications.europa.eu/resource/authority/corporate-body/CNECT	"Data archive containing files in the following formats: MS-Word doc"

https://data.europa.eu/euodp/en/linked-data

Schema.org

- Started by Google, Microsoft, Yahoo, and Yandex to help with indexing web pages for search
- Schema.org metadata can be embedded using microdata, RDFa or JSON-LD
- Commonly used types
  - Creative works: CreativeWork, Book, Movie, MusicRecording, Recipe, TVSeries
  - Embedded non-text objects: AudioObject, ImageObject, VideoObject
  - Event
  - Organization
  - Person
  - Place, LocalBusiness, Restaurant ...
  - Product, Offer, AggregateOffer
  - Review, AggregateRating

schema.org

Google Dataset Search

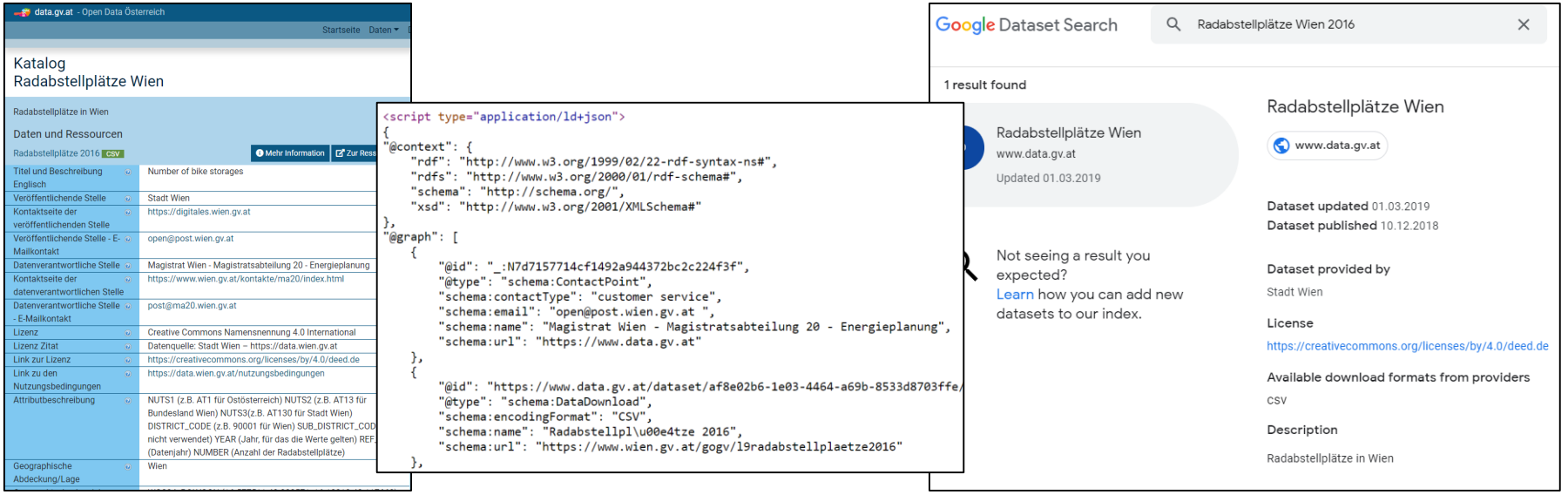
- schema.org:Dataset
  - based on W3C DCAT
- Full definition
  - https://schema.org/Dataset
- Google has an *application profile*
  - https://developers.google.com/search/docs/data-types/dataset

Required properties	
`description`	Text A short summary describing a dataset.
`name`	Text A descriptive name of a dataset. For example, "Snow depth in Northern Hemisphere".

Recommended properties

Schema.org - example

- View page source of any dataset at data.gv.at
  - Navigate to <script> section
- Search in Google Dataset Search for ‘Radabstellplätze Wien 2016’



data.gv.at - Open Data Österreich

Startseite Daten

Katalog
Radabstellplätze Wien

Radabstellplätze in Wien

Daten und Ressourcen

Radabstellplätze 2016 CSV

Mehr Information | Zur Ressource

Titel und Beschreibung Englisch	Number of bike storages
Veröffentlichende Stelle	Stadt Wien
Kontaktseite der veröffentlichenden Stelle	https://digitales.wien.gv.at
Veröffentlichende Stelle - E-Mailkontakt	open@post.wien.gv.at
Datenverantwortliche Stelle	Magistrat Wien - Magistratsabteilung 20 - Energieplanung
Kontaktseite der datenverantwortlichen Stelle	https://www.wien.gv.at/kontakte/ma20/index.html
Datenverantwortliche Stelle - E-Mailkontakt	post@ma20.wien.gv.at
Lizenz	Creative Commons Namensnennung 4.0 International
Lizenz Zitat	Datenquelle: Stadt Wien – https://data.wien.gv.at
Link zur Lizenz	https://creativecommons.org/licenses/by/4.0/deed.de
Link zu den Nutzungsbedingungen	https://data.wien.gv.at/nutzungsbedingungen
Attributbeschreibung	NUTS1 (z.B. AT1 für Ostösterreich) NUTS2 (z.B. AT13 für Bundesland Wien) NUTS3(z.B. AT130 für Stadt Wien) DISTRICT_CODE (z.B. 90001 für Wien) SUB_DISTRICT_CODE(0 da nicht verwendet) YEAR (Jahr, für das die Werte gelten) REF_YEAR (Datenjahr) NUMBER (Anzahl der Radabstellplätze)
Geographische Abdeckung/Lage	Wien

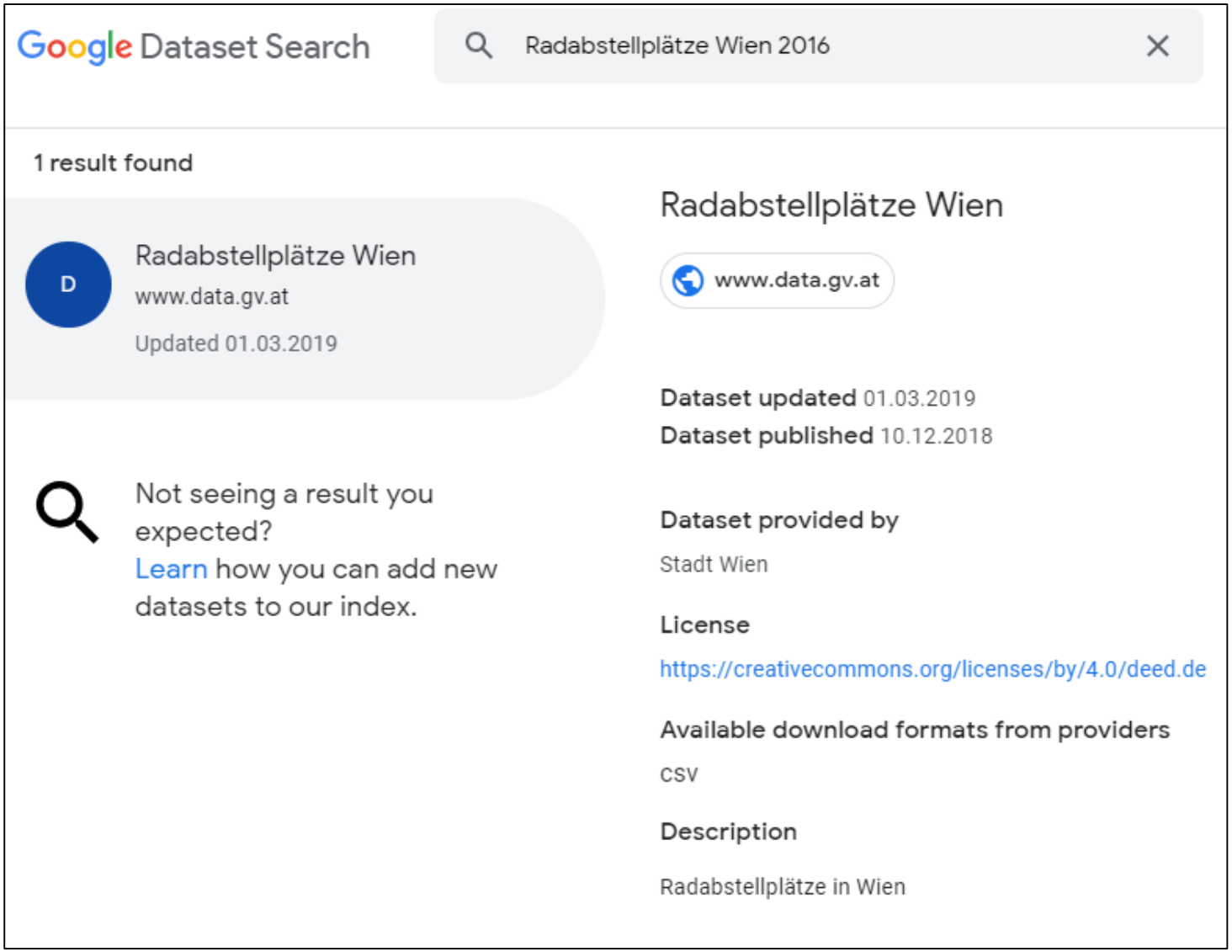
```
<script type="application/ld+json">
{
"@context": {
    "rdf": "http://www.w3.org/1999/02/22-rdf-syntax-ns#",
    "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
    "schema": "http://schema.org/",
    "xsd": "http://www.w3.org/2001/XMLSchema#"
},
"@graph": [
    {
        "@id": "_:N7d7157714cf1492a944372bc2c224f3f",
        "@type": "schema:ContactPoint",
        "schema:contactType": "customer service",
        "schema:email": "open@post.wien.gv.at ",
        "schema:name": "Magistrat Wien - Magistratsabteilung 20 - Energieplanung",
        "schema:url": "https://www.data.gv.at"
    },


   "@id": "https://www.data.gv.at/dataset/af8e02b6-1e03-4464-a69b-8533d8703ffe",
   "@type": "schema:Dataset",
   "schema:dateModified": "2019-03-01T10:20:42.981483",
   "schema:datePublished": "2018-12-10T14:46:53.400738",
   "schema:description": "Radabstellpl\u00e4tze in Wien\r\n",
   "schema:distribution": {
       "@id": "https://www.data.gv.at/dataset/af8e02b6-1e03-4464-a69b-8533d8703ffe/resour
   },
   "schema:includedInDataCatalog": {
       "@id": "_:N561 1dfe394574bbf9e3200474914654a"
```

Google Dataset Search Beta

Search for Datasets

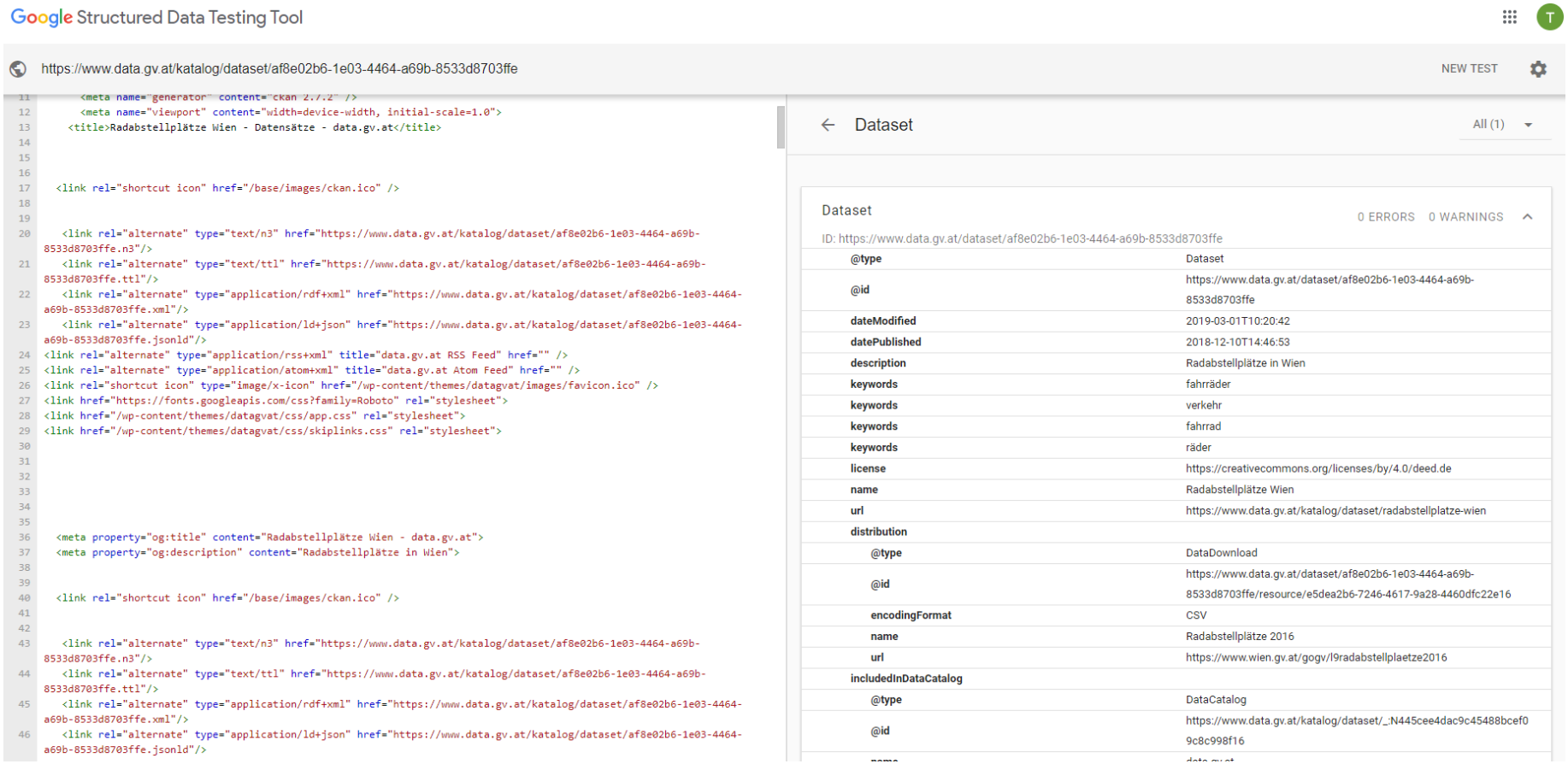
Try boston education data or weather site:noaa.gov

Learn more about including your datasets in Dataset Search.



https://toolbox.google.com/datasetsearch

Structured Data Testing Tool



https://search.google.com/structured-data/testing-tool

Structured Data Testing Tool

Dataset

0 ERRORS 0 WARNINGS

ID: https://www.data.gv.at/dataset/af8e02b6-1e03-4464-a69b-8533d8703ffe

@type	Dataset
@id	https://www.data.gv.at/dataset/af8e02b6-1e03-4464-a69b-8533d8703ffe
dateModified	2019-03-01T10:20:42
datePublished	2018-12-10T14:46:53
description	Radabstellplätze in Wien
keywords	fahrräder
keywords	verkehr
keywords	fahrrad
keywords	räder
license	https://creativecommons.org/licenses/by/4.0/deed.de
name	Radabstellplätze Wien
url	https://www.data.gv.at/katalog/dataset/radabstellplatze-wien
distribution	
@type	DataDownload
@id	https://www.data.gv.at/dataset/af8e02b6-1e03-4464-a69b-8533d8703ffe/resource/e5dea2b6-7246-4617-9a28-4460dfc22e16
encodingFormat	CSV
name	Radabstellplätze 2016
url	https://www.wien.gv.at/gogv/l9radabstellplaetze2016

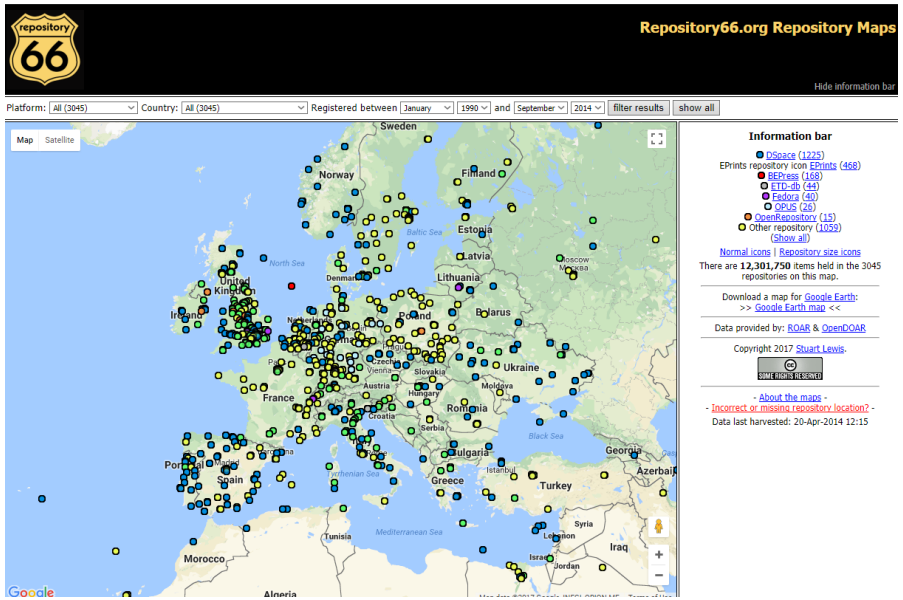
Synergy between DOIs and schema.org

- DataCite provides metadata on DOIs
  - RDFs compliant with schema.org
- Example
  - https://data.datacite.org/application/vnd.schemaorg.ld+json/**<DOI>**
  - https://data.datacite.org/application/vnd.schemaorg.ld+json/**10.1371/journal.pcbi.1006750**
- JSON-LD returned by DataCite can be directly embedded on web pages

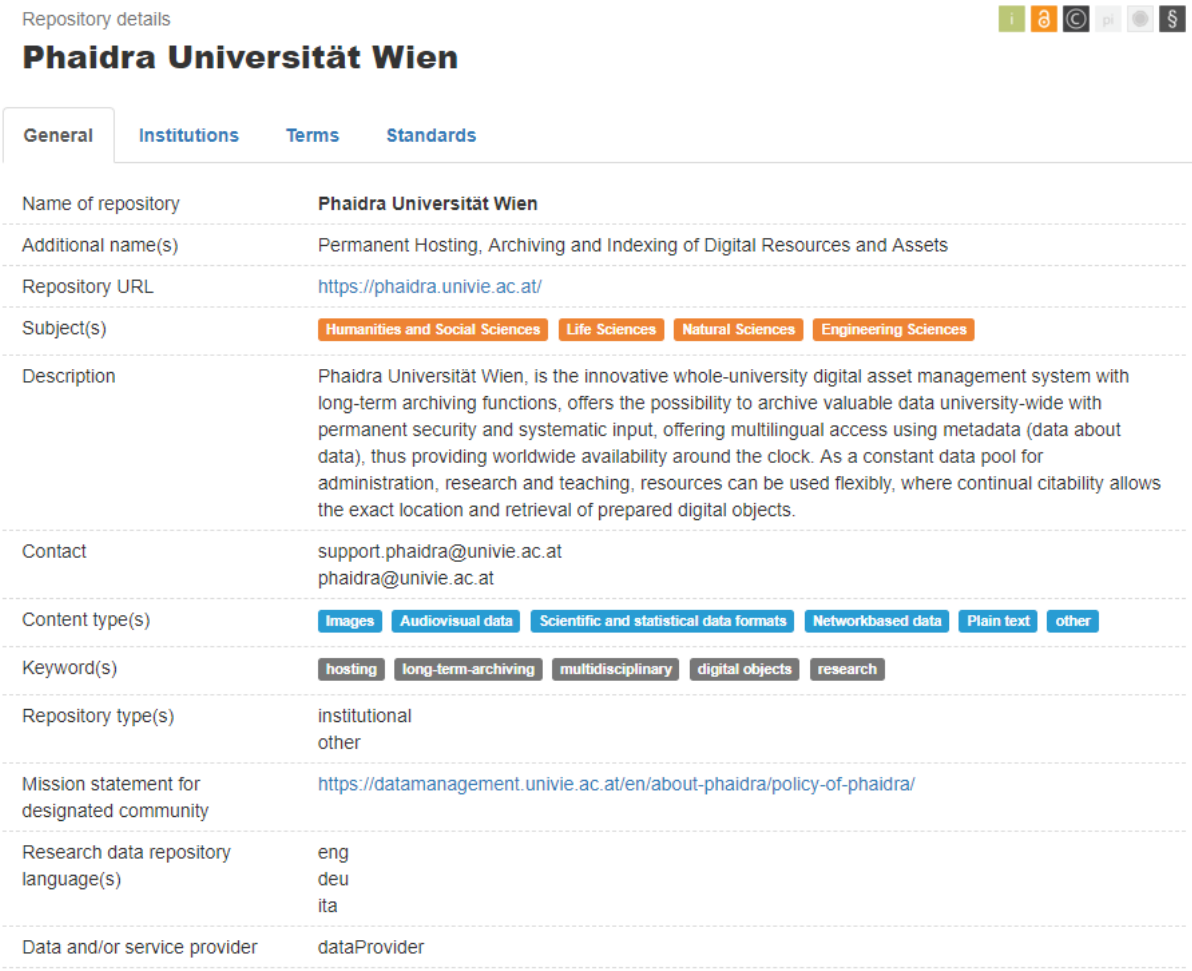
https://data.datacite.org

Repository registries

- Directory of Open Access Repositories – DOAR
  - Based on registrations
  - http://www.opendoar.org/
- Registry of Open Access Repositories – ROAR
  - Automatically harvested list based on OAI-PMH
  - http://roar.eprints.org/
- Projection of DOAR and ROAR on maps
  - http://maps.repository66.org
- re3data.org



Repository registries – re3data



Repository details

Phaidra Universität Wien

General | Institutions | Terms | Standards

Name of repository	**Phaidra Universität Wien**
Additional name(s)	Permanent Hosting, Archiving and Indexing of Digital Resources and Assets
Repository URL	https://phaidra.univie.ac.at/
Subject(s)	Humanities and Social Sciences; Life Sciences; Natural Sciences; Engineering Sciences
Description	Phaidra Universität Wien, is the innovative whole-university digital asset management system with long-term archiving functions, offers the possibility to archive valuable data university-wide with permanent security and systematic input, offering multilingual access using metadata (data about data), thus providing worldwide availability around the clock. As a constant data pool for administration, research and teaching, resources can be used flexibly, where continual citability allows the exact location and retrieval of prepared digital objects.
Contact	support.phaidra@univie.ac.at phaidra@univie.ac.at
Content type(s)	Images; Audiovisual data; Scientific and statistical data formats; Networkbased data; Plain text; other
Keyword(s)	hosting; long-term-archiving; multidisciplinary; digital objects; research
Repository type(s)	institutional other
Mission statement for designated community	https://datamanagement.univie.ac.at/en/about-phaidra/policy-of-phaidra/
Research data repository language(s)	eng deu ita
Data and/or service provider	dataProvider

https://www.re3data.org/repository/r3d100010472

External visibility - summary

- DOI Registration bodies (DataCite)
  - Metadata schema
- OAI-PMH
  - architecture
- OpenAIRE
- DCAT
  - Metadata schema
- Schema.org and Google Dataset Search
- Repository registries

CONCLUSION

You should know

- what is a repository system?
- how to compare repository systems?
- how to make repository contents visible?
  - What are the options and how to choose the best one?

Next lecture

- 20.05
- Developing Research Data Management Services
  - Data lifecycle model
  - Policies
  - Costs and Business Models
  - Data Stewards
  - Repository Certification

References

- A comparison of research data management platforms: architecture, flexible metadata and interoperability
  - https://doi.org/10.1007/978-3-319-16486-1
- Institutional Repository Software Comparison
  - https://works.bepress.com/jean_gabriel_bankier/22/
- Open Source Software for Digital Preservation Repositories: a Survey
  - https://arxiv.org/abs/1707.06336
- Research Data Repositories: Review of current features, gap analysis, and recommendations for minimum requirements
  - https://www.rdc-drc.ca/wp-content/uploads/Review-of-Research-Data-Repositories-2015.pdf
- Institutional repository software comparison: DSpace, EPrints, Digital Commons, Islandora and Hydra
  - https://dx.doi.org/10.14288/1.0075768