Research Data Management

Open Access, Open Data

Data Management Plan

Yuri Demchenko, EDISON
University of Amsterdam

13 July 2017
SNE Group meeting, UvA, Amsterdam

EDISON
building the data science profession

EDISON – **E**ducation for **D**ata **I**ntensive **S**cience to **O**pen **N**ew science frontiers

Grant 675419 (INFRASUPP-4-2015: CSA)

Outline

- Background: Open Access, Open Data, Open Science
  - CODATA, RDA, others
  - H2020 Programs and projects: OpenAIRE, FOSTER
  - FAIR Initiative and Data Stewardship
- Services for research community
  - OpenAIRE, Zenodo
  - PID, ORCID
- Data Management Plan (DMP) in/by UvA
- Research Data Management training

Open Access to Scientific Publications

- EC initiative on Open Access scientific publications from publicly funded projects
  - Included into Declaration from the H2020 Rome meeting (2012)
  - Approx 3500 publicly funded ROs and 2000 privately funded ROs
  - Special funding scheme for reimbursing publications
  - Issues with China, India, Russia compliance to OA principles
    - Consultation at high governmental level
- OpenAIRE project is exploring models for open access to publications
  - PID (Persistent ID for data), ORCHID (Open Researcher ID), Linked data
  - Zenodo – spinoff service for Open Access of research data/information
- Community initiative - Panton Principles for Open Data in Science (http://pantonprinciples.org/)

Open Research Data

- Research data can be defined as whatever is either produced in the research process or evidences research outputs such as articles
- The European Commission’s Research Data definition is: *“information, in particular facts or numbers, collected to be examined and considered and as a basis for reasoning, discussion, or calculation”*
  - http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf
- Examples include: statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings, images
- Open data are deposited in institutional or specialist repositories and licensed appropriately so that prospective users know clearly any limitations on re-use.

Open and Toll Access (OA and TA)

- Open Access generally refers to the outputs of research, such as journal articles, as distinct from research data, which are produced as part of the research process
- Open Access is differentiated from the traditional method of access to research outputs, known as Toll Access
  - Toll Access can be by means of institutional or personal subscription to journals, or to aggregations of content, or by means of paying publishers for access to individual articles
  - Toll Access payment is reader-side

Open Access Definition

Budapest Open Access Initiative (BOAI) 2002, reaffirmed in 2012:

- By "open access" to … literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited.
  - http://www.budapestopenaccessinitiative.org/boai-10-recommendations

Peter Suber’s Concise Definition:

- Open Access literature is “digital, online, free of charge, and free of most copyright and licensing restrictions” (Suber, P. Open access. MIT Press, 2012. Available at:
  - https://mitpress.mit.edu/sites/default/files/9780262517638_Open_Access_PDF_Version.pdf
- FAIR principles by DTLS.nl (Barend Mons) and EU H2020

Gratis and Libre OA

- Context:
  Intellectual property laws generally offer limited “fair dealing” or “fair use” exemptions
- Gratis OA is free of charge to access but subject to the limits of fair dealing
  - it removes toll barriers but not permission barriers
- Libre OA is both free of charge and free of at least some legal and licensing restrictions
  - it removes toll barriers and at least some permission barriers
- The BOAI definition is Libre.

Green OA –1 and Green OA - 2

Green OA -1 is delivered through **self-archiving**: authors deposit manuscripts in institutional or disciplinary repositories;

- Relies on a recent but well established infrastructure of repositories
- Is easy and cheap: each article only incurs a very small portion of the overhead costs of setting up and running repositories
- Does not incur the overheads of peer-review;
- However, deposited articles may be, most often have been, peer-reviewed for publication in traditional Toll Access journals

Green OA – 2 is compatible with subscription journal publishing: scholars can publish in TA journals and, through self-archiving, still make their articles OA

- Is often subject to an **embargo period** imposed by publishers, generally of between 6 and 12 months
- Depends on authors' obtaining rights from publishers to deposit and make articles available
- Is hospitable to many other types of document, notably pre-prints, theses, and reports.

Gold OA-1 and Gold OA-2

Gold OA – 1: Offers articles that are **paid for by the authors or their institutions or funders**

- Articles may be either in completely OA journals or in hybrid journals, containing both OA and TA articles
- Articles are peer-reviewed for publication
- Incurs much the same costs for the editorial and peer review process as TA journal publishing
- Is always immediate, while Green OA is often subject to time embargoes imposed by subscription journal publishers.

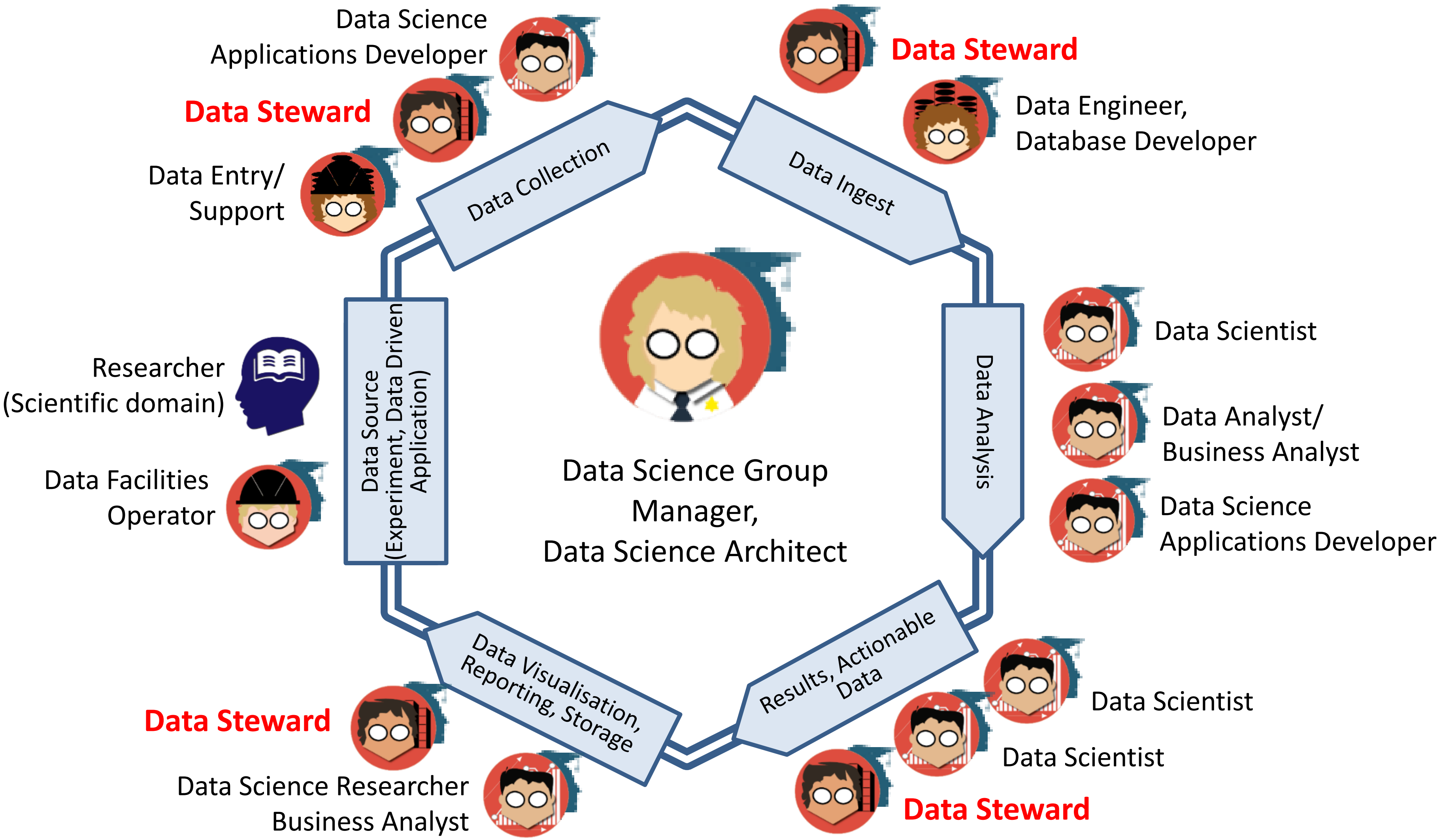
Gold OA – 2: Provides access to the published version of an article, while Green OA generally provides **access only to the author's final peer-reviewed manuscript**, without the formatting or pagination of the published version

- By its nature is confined to post-prints
- Generally obtains rights and permissions direct from the rights-holder (usually the author);
- Is delivered through journals: these may be completely OA or hybrid, where some articles are OA and others toll access;
- Both Green and Gold OA are gratis. Green OA generally is only gratis; Gold OA may be Libre.

GO FAIR and IFDS

- Global Open FAIR
  - Findable – Accessible – Interoperable - Reusable
- IFDS – Internet of FAIR Data and Services = EOSC
- GO FAIR implementation approach
  - GO-BUILD
  - GO-CHANGE
  - GO-TRAIN: Training of data stewards capable of providing FAIR data services
- A critical success factor is availability of expertise in **Data Stewardship**
  - Training of a new generation of FAIR data experts is urgently needed to provide the necessary capacity.

Building a Data Science Team


Data Science Applications Developer
Data Steward
Data Entry/ Support
Data Steward
Data Engineer, Database Developer
Data Collection
Data Ingest
Researcher (Scientific domain)
Data Facilities Operator
Data Source (Experiment, Data Driven Application)
Data Science Group Manager, Data Science Architect
Data Analysis
Data Scientist
Data Analyst/ Business Analyst
Data Science Applications Developer
Data Visualisation, Reporting, Storage
Results, Actionable Data
Data Steward
Data Science Researcher Business Analyst
Data Scientist
Data Scientist
Data Steward

Data Science or Data Management Group/Department: Organisational structure and staffing - EXAMPLE

Data Science or Data Management Group/Department

>> Reporting to CDO/CTO/CEO

- Providing cross-organizational services

- (Managing) Data Science Architect (1)
- Data Scientist (1), Data Analyst (1)
- Data Science Application programmer (2)
- Data Infrastructure/facilities administrator/operator: storage, cloud, computing (1)
- **Data stewards**, curators, archivists (3-5)

Estimated: Group of 10-12 data specialists for research institution of 200-300 research staff.

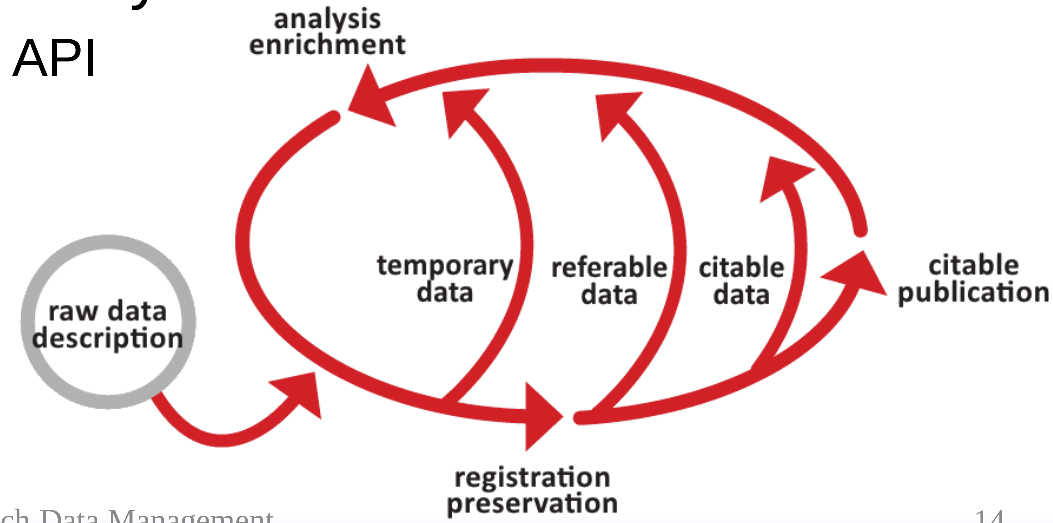
Growing role and demand for Data Stewards and data stewardship

Data Stewards – A rising new role in Data Science ecosystem

- Data Stewards as a key bridging role between Data Scientists as (hard)core data experts and scientific domain researchers (HLEG EOSC report)
- Current definition of Data Steward (part of Data Science Professional profiles)
  - Data Steward is a data handling and management professional whose responsibilities include planning, implementing and managing (research) data input, storage, search, and presentation.
  - Data Steward creates data model for domain specific data, support and advice domain scientists/ researchers during the whole research cycle and data management lifecycle.

Persistent Identifier (PID)

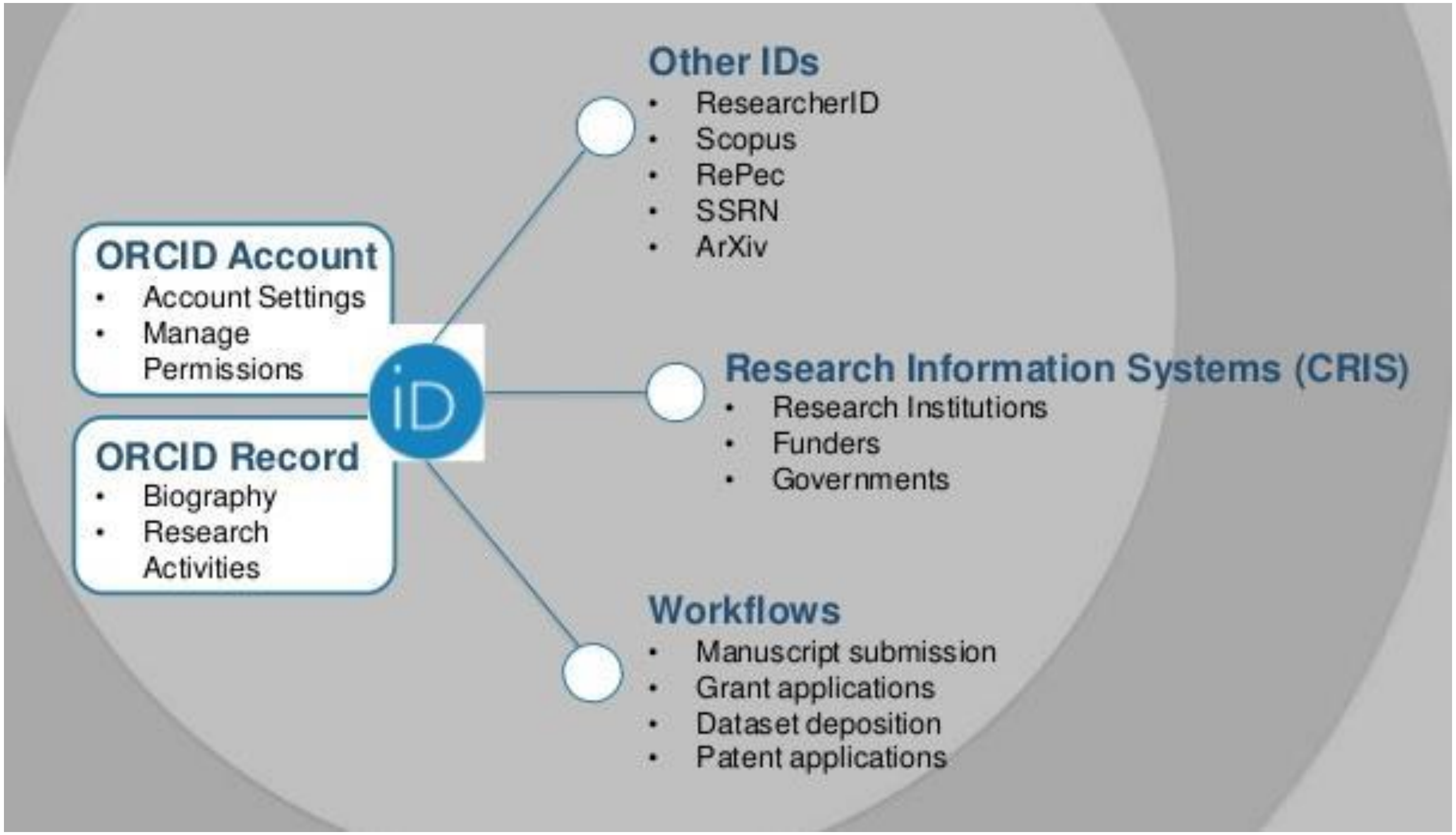
- PID – Persistent Identifier for Digital Objects
  - Managed by European PID Consortium (EPIC) http://www.pidconsortium.eu/
  - Superset of DOI - Digital Object Identifier (http://www.doi.org/)
  - Handle System by CNRI (Corporation for National Research Initiatives) for resolving DOI (http://www.handle.net/)
- PID provides a mechanism to link data during the whole research data transformation cycle
  - EPIC RESTful Web Service API published May 2013



ORCID - Connecting research and researchers

- Research in the digital realm is becoming increasingly linked up
  - Leverage this to increase your profile
  - Get an **ORCID (Open Researcher and Contributor ID)** and identify yourself as a unique researcher
  - ORCID provides a persistent digital identifier that distinguishes you from every other researcher i.e. that Dr. John Smith
  - Looks something like: http://orcid.org/xxxx-xxxx-xxxx-xxxx
  - Simple and free to register at: http://orcid.org/

Connecting research and researchers



ORCID Account
- Account Settings
- Manage Permissions

ORCID Record
- Biography
- Research Activities

Other IDs
- ResearcherID
- Scopus
- RePec
- SSRN
- ArXiv

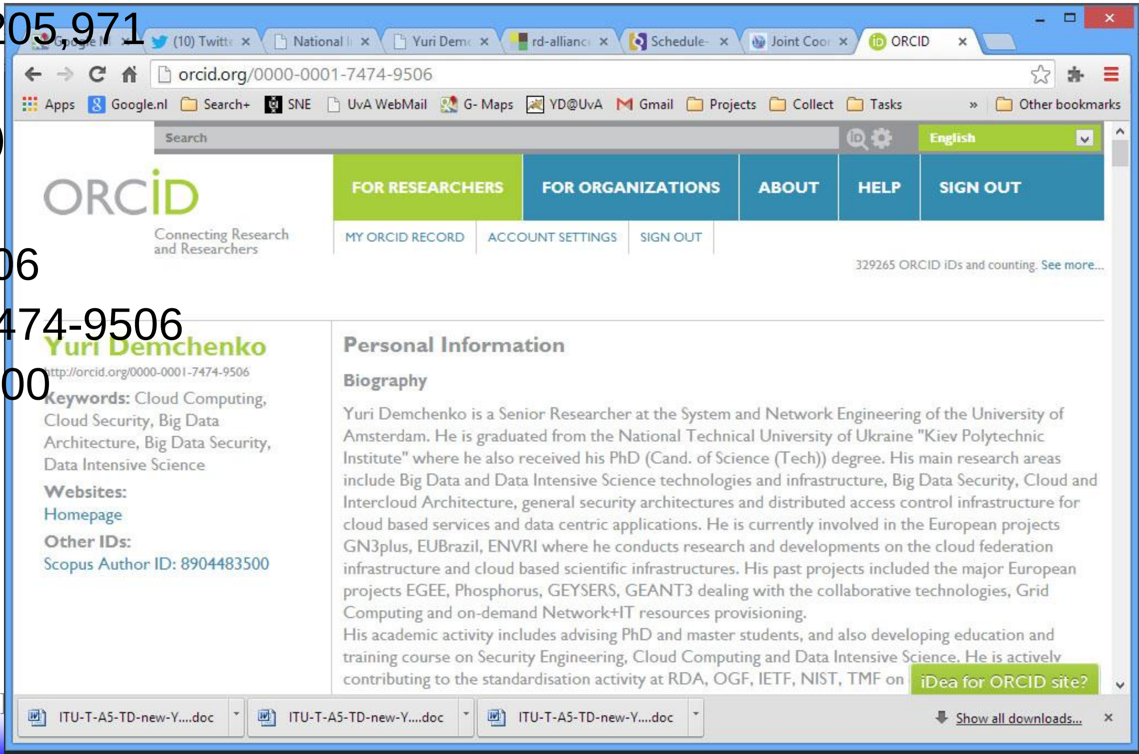
Research Information Systems (CRIS)
- Research Institutions
- Funders
- Governments

Workflows
- Manuscript submission
- Grant applications
- Dataset deposition
- Patent applications

- Link together your research
- Source: ORCID: Connecting Research and Researchers,
- Biblioteca del Campus Terrassa on Jul 11, 2013

ORCID (Open Researcher and Contributor ID)

- ORCID is a nonproprietary alphanumeric code to uniquely identify scientific and other academic authors
  - Launched October 2012
- ORCID Statistics – July 2017 **3,603,217**
  - Live ORCID IDs May 2014 - 511, 203 (October 2013 - 329,265)
  - ORCID IDs with at least one work **1,381,172** (May (2014 - 121,529; October 2013 - 79,332)
  - Works **22,056,406** (2014 - 205,971
  - Works with unique DOIs **9,306,782 2014** - 1,267,083)
- Personal ORCID
  - ORCID 0000-0001-7474-9506
  - http://orcid.org/0000-0001-7474-9506
  - Scopus Author ID 8904483500



329265 ORCID iDs and counting. See more...

Yuri Demchenko

http://orcid.org/0000-0001-7474-9506

Keywords: Cloud Computing, Cloud Security, Big Data Architecture, Big Data Security, Data Intensive Science

Websites:

Homepage

Other IDs:

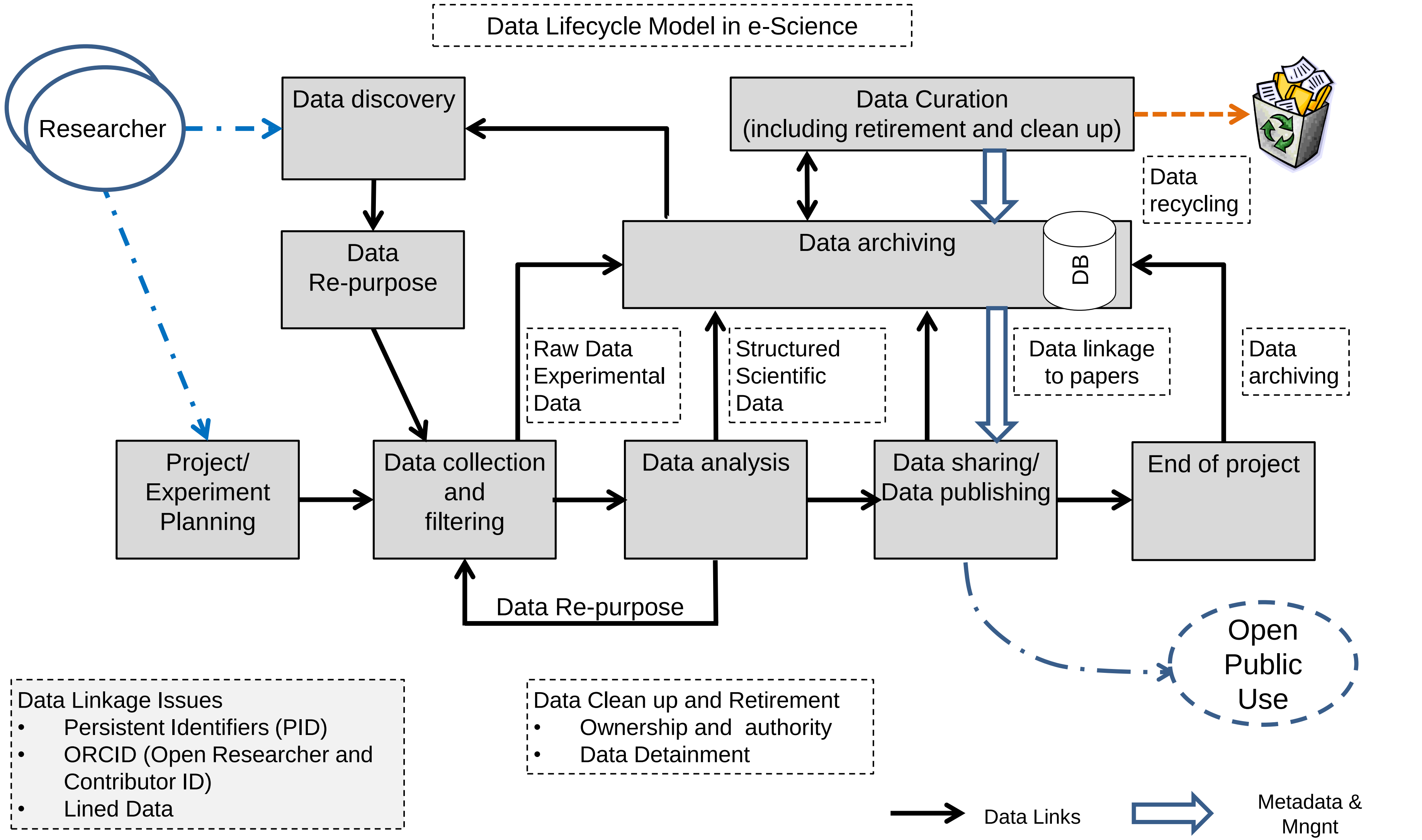
Scopus Author ID: 8904483500

Personal Information

Biography

Yuri Demchenko is a Senior Researcher at the System and Network Engineering of the University of Amsterdam. He is graduated from the National Technical University of Ukraine "Kiev Polytechnic Institute" where he also received his PhD (Cand. of Science (Tech)) degree. His main research areas include Big Data and Data Intensive Science technologies and infrastructure, Big Data Security, Cloud and Intercloud Architecture, general security architectures and distributed access control infrastructure for cloud based services and data centric applications. He is currently involved in the European projects GN3plus, EUBrazil, ENVRI where he conducts research and developments on the cloud federation infrastructure and cloud based scientific infrastructures. His past projects included the major European projects EGEE, Phosphorus, GEYSERS, GEANT3 dealing with the collaborative technologies, Grid Computing and on-demand Network+IT resources provisioning.
His academic activity includes advising PhD and master students, and also developing education and training course on Security Engineering, Cloud Computing and Data Intensive Science. He is actively contributing to the standardisation activity at RDA, OGF, IETF, NIST, TMF on

Scientific Data Lifecycle Model



Data Lifecycle Model in e-Science

Researcher

Data discovery

Data Re-purpose

Data Curation (including retirement and clean up)

Data recycling

Data archiving

DB

Raw Data Experimental Data

Structured Scientific Data

Data linkage to papers

Data archiving

Project/ Experiment Planning

Data collection and filtering

Data analysis

Data sharing/ Data publishing

End of project

Data Re-purpose

Open Public Use

Data Linkage Issues
- Persistent Identifiers (PID)
- ORCID (Open Researcher and Contributor ID)
- Lined Data

Data Clean up and Retirement
- Ownership and authority
- Data Detainment

Data Links

Metadata & Mngnt

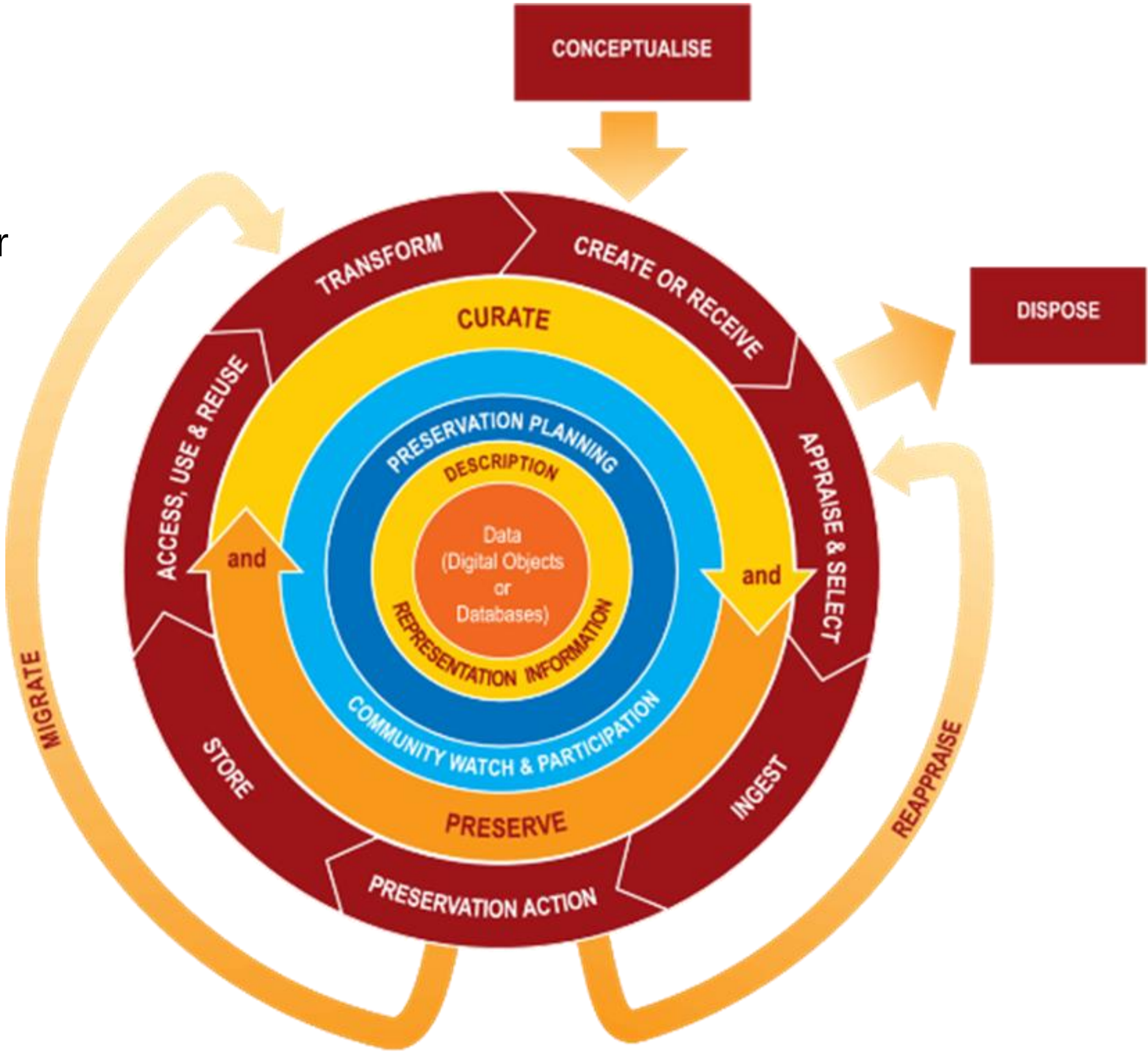
DCC Curation Lifecycle Model - Actions

Three sets of actions:

- Sequential Actions (7+1): key actions needed as data move through their lifecycle
- Occasional Actions (3): only occur when special conditions are met, but they do not apply to all data
- Full Lifecycle Actions (4): apply to all stages in the lifecycle

The DCC Curation Lifecycle Model is based on the OAIS Reference Model

- OAIS = Open Archival Information System (pictured)
- OAIS is a model that defines a generic framework for building a digital archive



What is a Data Management Plan?

A brief plan written at the start of a project to define:

- What data will be collected or created?
- How the data will be documented and described?
- Where the data will be stored?
- Who will be responsible for data security and backup?
- Which data will be shared and/or preserved?
- How the data will be shared and with whom?

Why develop a DMP?

DMPs are often submitted with grant applications, but are useful whenever researchers are creating data.

They can help researchers to:

- Make informed decisions to anticipate & avoid problems.
- Develop procedures early on for consistency.
- Ensure data are accurate, complete, reliable and secure.
- Avoid duplication, data loss and security breaches.
- Save time and effort to make their lives easier!

Themes to address in DMPs

- Data collection
- Documentation and metadata
- Ethics and legal compliance
- Storage and backup
- Selection and preservation
- Data sharing
- Responsibilities and resources

- Roadmap for Research Data from the League of European Research Universities (LERU)
  - http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf
- Other examples and tools
  - https://dmponline.dcc.ac.uk
  - https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/dmp/framework.html

Data licensing

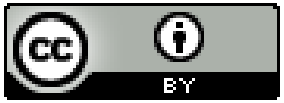
- Creative Commons licenses are not always suitable for data because data have different IPR than generic digital content
  - http://creativecommons.org/licenses/

cc BY
cc BY SA
cc BY ND
cc BY NC
cc BY NC ND
cc BY NC SA

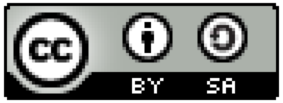
- Open Data Commons have specific licenses for data that conform to the Open Knowledge Foundation's definition of Open Data: http://opendatacommons.org/guide/
- They have 2 basic options:
  - Public Domain: puts all the materials in the public domain
  - Share-Alike (plus Attribution): similar to the Creative Commons Attribution Share-Alike license
- Also see the following guides:
  - http://datalib.edina.ac.uk/mantra/preservation.html: an online learning unit from Mantra on "Sharing, preservation, and licensing", go to slides 15-17
  - http://infteam.jiscinvolve.org/wp/2012/10/09/opendatalicensing: Open Data licensing animation video
  - http://opendefinition.org/guide/data/

Data licensing

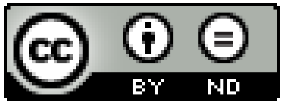
- Creative Commons licenses are not always suitable for data because data have different IPR than generic digital content
  - http://creativecommons.org/licenses/



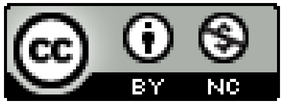
Attribution
CC BY



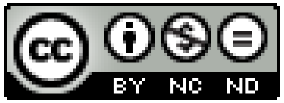
Attribution-ShareAlike
CC BY-SA



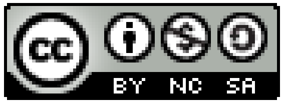
Attribution-NoDerivs
CC BY-ND



Attribution-NonCommercial
CC BY-NC



Attribution-NonCommercial-ShareAlike
CC BY-NC-SA



Attribution-NonCommercial-NoDerivs
CC BY-NC-ND

Practical Example

- Data Management Plan checklist
  - University of Amsterdam DMP template

KAG3-DSDM: *Data Management group: data curation, preservation and data infrastructure*

DM-BoK version 2 “Guide for performing data management” – 11 Knowledge Areas

- (1) Data Governance
- (2) Data Architecture
- (3) Data Modelling and Design
- (4) Data Storage and Operations
- *(5) Data Security*
- (6) Data Integration and Interoperability
- *(7) Documents and Content*
- (8) Reference and Master Data
- (9) Data Warehousing and Business Intelligence
- ***(10) Metadata***
- (11) Data Quality

Other Knowledge Areas motivated by RDA, European Open Data initiatives, European Open Data Cloud

- (12) PID, metadata, data registries
- (13) Data Management Plan
- (14) Open Science, Open Data, Open Access, ORCID
- (15) Responsible data use

- Highlighted in red: Considered (Research) Data Management literacy (minimum required knowledge)

Discussion

Useful links

- EDISON project website http://edison-project.eu/

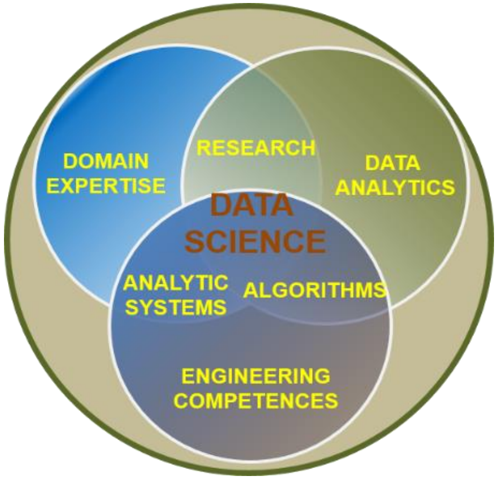
- EDISON Data Science Framework Release 1 (EDSF) http://edison-project.eu/edison-data-science-framework-edsf
  - Data Science Competence Framework http://edison-project.eu/data-science-competence-framework-cf-ds
  - Data Science Body of Knowledge http://edison-project.eu/data-science-body-knowledge-ds-bok
  - Data Science Model Curriculum http://edison-project.eu/data-science-model-curriculum-mc-ds
  - Data Science Professional Profiles http://edison-project.eu/data-science-professional-profiles-definition-dsp

- Survey Data Science Competences: Invitation to participate https://www.surveymonkey.com/r/EDISON_project_-_Defining_Data_science_profession

Data Scientist definition

Based on the definitions by NIST Big Data WG (NIST SP1500 - 2015)

- *A **Data Scientist** is a practitioner who has sufficient knowledge in the overlapping regimes of expertise in **business needs, domain knowledge, analytical skills, and programming and systems engineering expertise** to manage the end-to-end scientific method process through each stage in the **big data lifecycle***
  - *… Till the delivery of an **expected scientific and business value** to science or industry*

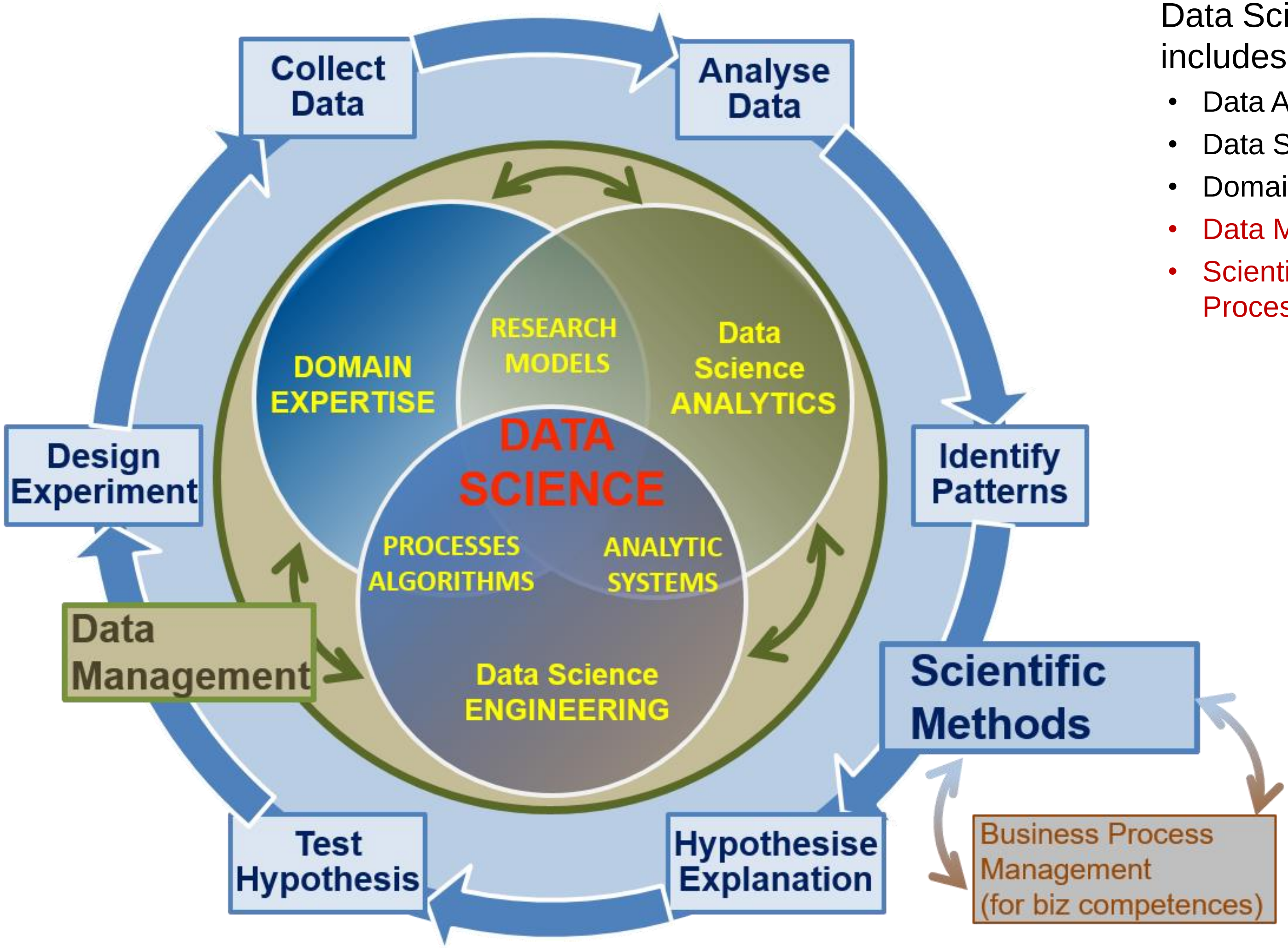


[ref] Legacy: NIST BDWG definition of Data Science

- *Other definitions to admit such features as*
  - Ability to solve variety of business problems
  - Optimize performance and suggest new services for the organisation
  - Develop a special mindset and be statistically minded, ***understand raw data** and **"appreciate data as a first class product"***

- ***Data science** is the empirical synthesis of actionable knowledge and technologies required to handle data from raw data through the complete data lifecycle process.*
- ***Big Data** is the technology to build system and infrastructures to process large volume of structurally complex data in a time effective way*

Data Science Competence Groups - Research



Data Science Competence includes 5 areas/groups

- Data Analytics
- Data Science Engineering
- Domain Expertise
- Data Management
- Scientific Methods (or Business Process Management)

Scientific Methods

- Design Experiment
- Collect Data
- Analyse Data
- Identify Patterns
- Hypothesise Explanation
- Test Hypothesis

Business Operations

- Operations Strategy
- Plan
- Design & Deploy
- Monitor & Control
- Improve & Re-design