Lecture 3 – Open Data

Advanced Data Management

Data Science and Scientific Computing / UniTS – DMG

Scientific and Data-Intensive Computing / UniTS – DMG

Data and Open Data

- Data and information are often used interchangeably, but relative to today's computers and transmission media, data is information converted into binary digital form
- Internet changed the way we use, share and access data
- These new technologies of sharing data provides unprecedented tools to get huge amounts of data to analyze
- Data can be private (e.g. restricted user access) or public
- “Public” in some context can be synonym of “open”, but there is not complete agreement on “open” definition
- There are several initiatives and consortiums that promote open data:
  - World Wide Web Consortium (W3C)
  - Open Society Foundations (OSF)
  - Open Knowledge International (OKI), formerly Open Knowledge Foundation (OKF)
  - Open Science as a Practice (openscienceASAP)
  - Research Data Alliance
  - … and many others

What does “open” mean?

- *The Open Definition* “Open means anyone can freely access, use, modify, and share for any purpose subject, at most, to requirements that preserve provenance and openness.” - Open Knowledge International
- *BOAI (Budapest Open Access Initiative)* “By 'open access' to this 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.” - Open Society Foundations
- Several fields where openness is applied, but they do not rely on the same exact definition of open:
  - Open Government
  - Open Source
  - Open Content
  - Open Science and Open Data Science
  - … and many others

Open Government

- Open Government is the governing doctrine which holds that citizens have the right to access the documents and proceedings of the government to allow for effective public oversight.
- There is a large number of areas where open government data are creating value, such as:
  - Transparency and democratic control
  - Participation
  - Improved efficiency and effectiveness of government services

Open Source

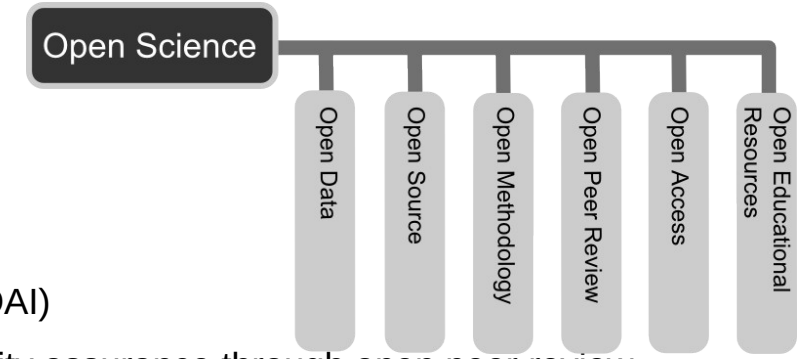
- The Open Source Definition is a document published by the Open Source Initiative, to determine whether a software license can be labeled with the open-source certification mark.
- The distribution terms of open-source software must comply some criteria, such as:
  - Software must include Source Code
  - Free Redistribution
  - The license must allow modifications and derived works, and must allow them to be distributed under the same terms as the license of the original software.
  - No Discrimination Against Persons or Groups
  - No Discrimination Against Fields of Endeavor
  - Distribution of License The rights attached to the program must apply to all to whom the program is redistributed without the need for execution of an additional license by those parties.
  - License Must Not Be Specific to a Product The rights attached to the program must not depend on the program's being part of a particular software distribution.
  - License Must Not Restrict Other Software The license must not place restrictions on other software that is distributed along with the licensed software.
  - License Must Be Technology-Neutral

Open Content

- Open content, coined by analogy with "open source" describes any kind of creative work including articles, pictures, audio, and video that is published in a format that explicitly allows the copying and the editing of the information
- The term applies to copyrightable content that is made freely available and licensed according to permission for what are known as the 5R activities:
  - Retain: Users may freely download, copy, store and manage the content.
  - Reuse: The content may be reused freely, for example on a website or in a class or workshop.
  - Revise: It is lawful to make changes to the content itself, for example reformatting or translating it.
  - Remix: The content may be combined in a mashup with other open content.
  - Redistribute: The content may be freely shared either in its original form or after being subjected to any permitted alteration.
- A Creative Commons (CC) license is one of several public copyright licenses that enable the free distribution of an otherwise copyrighted work

Open Science and Open Science Data

- “Open science is the idea that scientific knowledge of all kinds should be openly shared as early as is practical in the discovery process.” - *Michael Nielsen OpenScience ASAP*
- 6 principles:
  - Open Methodology: Document methods used
  - Open Source: Use open source technology
  - Open Data: Make available data freely available
  - Open Access: Data accessible to everyone (see BOAI)
  - Open Peer Review: Transparent and traceable quality assurance through open peer review
  - Open Educational Resources: Use Free and Open Materials for Education and University Teaching



- Open Science Data focuses on publishing observations and results of scientific activities available for anyone to analyze and reuse
  - Allow the verification of scientific claims
  - Allow data discovery from many sources to be integrated to give new knowledge
- Citations, references and acknowledgments are used to measure both the productivity and citation impact of the publications of a scientist
  - Conflict between the desire of scientists to have access to shared resources versus the desire of individual entities to profit when other entities partake of their resources

Open Data

- Different fields and communities have different needs about Open Data:
  - Transparency
  - Accessibility
  - Re-usability
  - Sharing
  - Knowledge distribution
  - Licensing issues
  - Citations and Acknowledgments
- Can we converge on a common definition of Open Data?

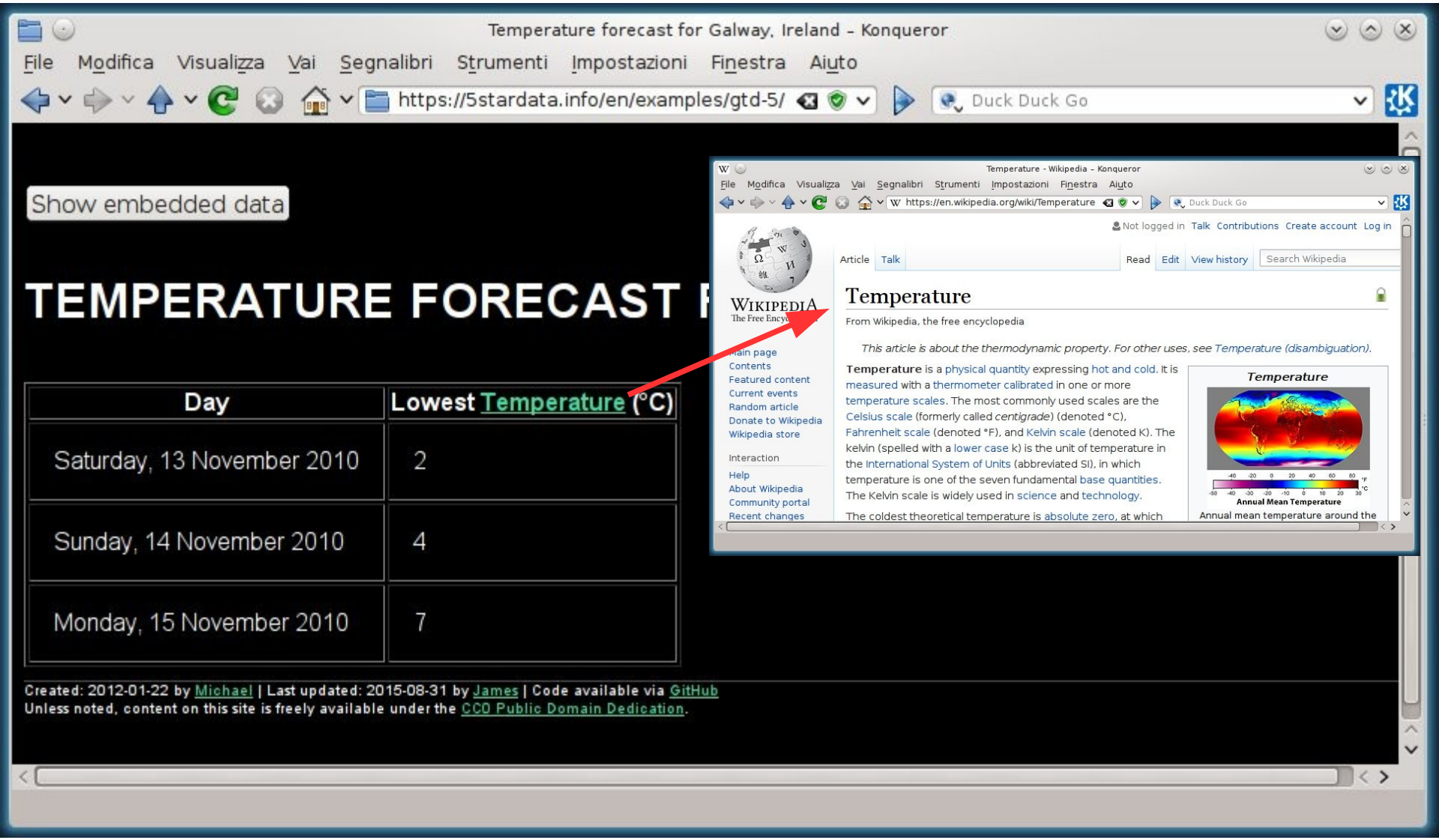
5 ★ Open Data

- Tim Berners-Lee, founder of the WorldWideWeb Consortium (W3C), suggested a 5-star deployment scheme for Open Data

★	make your stuff available on the Web (whatever format) under an open license
★★	make it available as structured data (e.g., Excel instead of image scan of a table)
★★★	make it available in a non-proprietary open format (e.g., CSV instead of Excel)
★★★★	use URIs to denote things, so that people can point at your stuff
★★★★★	link your data to other data to provide context

https://5stardata.info/en/

Example – Temperature Forecast



★★★★ data available through a URL

Linked Data (1)

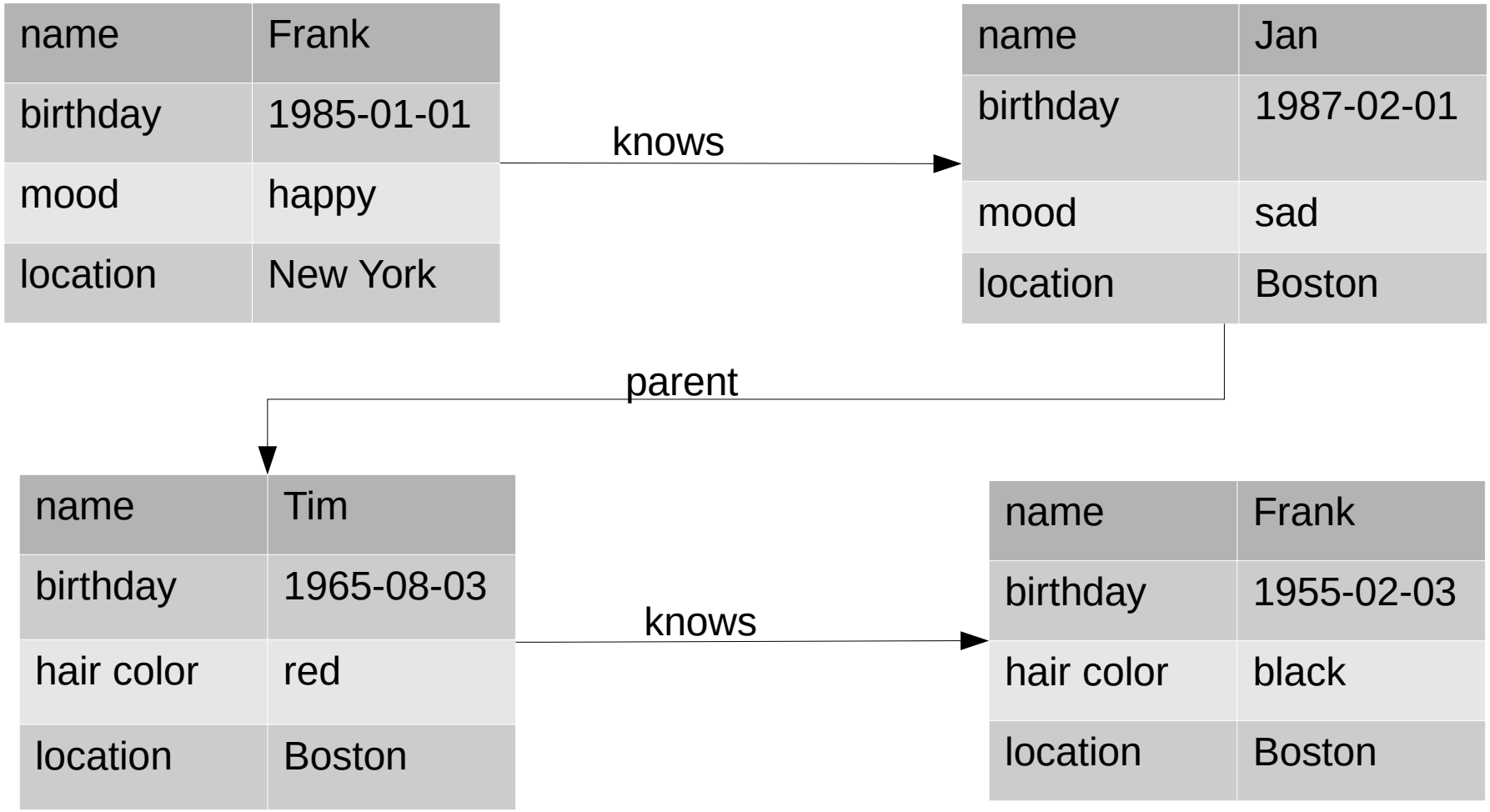
- There are many different types of Open Data we can use by itself: images, documents, videos, websites, etc.
- The web allows to make available images inside text, to mix different data together through link from one document to another
- Open Data can be combined into something more interesting than original pieces
- Computer does not really understand what links are about as an human
  - We need to deconstruct information and package it in a way understandable to a computer
  - We need to link this information it in a way understandable to a computer

Linked Data (2)

name	Frank
birthday	1985-01-01
mood	happy
location	New York

- Data are in a standard and usable format
- Common formats for data:
  - XML
  - Json
  - CSV
  - RDF
- How “Frank” is related with the rest of the world?

Linked Data (3)



- Relations link “Frank” with the rest of the world

Linked Data (4)

http://mysite.com/frank

name	Frank
birthday	1985-01-01
mood	happy
location	New York

http://schema.org/knows

http://myweb.com/jan

name	Jan
birthday	1987-02-01
mood	sad
location	Boston

http://schema.org/parent

http://othersite.com/tim

name	Tim
birthday	1965-08-03
hair color	red
location	Boston

http://schema.org/knows

http://othersite.com/frank

name	Frank
birthday	1955-02-03
hair color	black
location	Boston

- URL can specify which “Frank”
- URL can define relations

Linked Data (5)

http://mysite.com/frank

name	Frank
birthday	1985-01-01
mood	happy
location	New York

http://schema.org/knows

http://myweb.com/jan

name	Jan
birthday	1987-02-01
mood	sad
location	Boston

http://schema.org/lives

http://towns.com/newyork

name	New York
latitude	40°43'N
longitude	74°00'W
state	New York

http://schema.org/lives

http://towns.com/boston

name	Boston
latitude	41°21'N
longitude	71°03'W
state	Massachusetts

- You can mix information from different vocabularies

Linked Data (6)

- Three rules for Linked Data
  - Data are in a standard and usable format
  - Use unique identifiers for Data
  - Data are relationships
- It is not a top-down system: linked data does not need everybody agrees on all the terms
- When anybody communicates any kind of information, uses a mix of different vocabularies
- You cherry pick information from different set of terms from different vocabularies (ontologies)

Example – Potato Chips

English Vocabulary

Japanese Vocabulary

Nutrition Facts Vocabulary

Food Industry Vocabulary

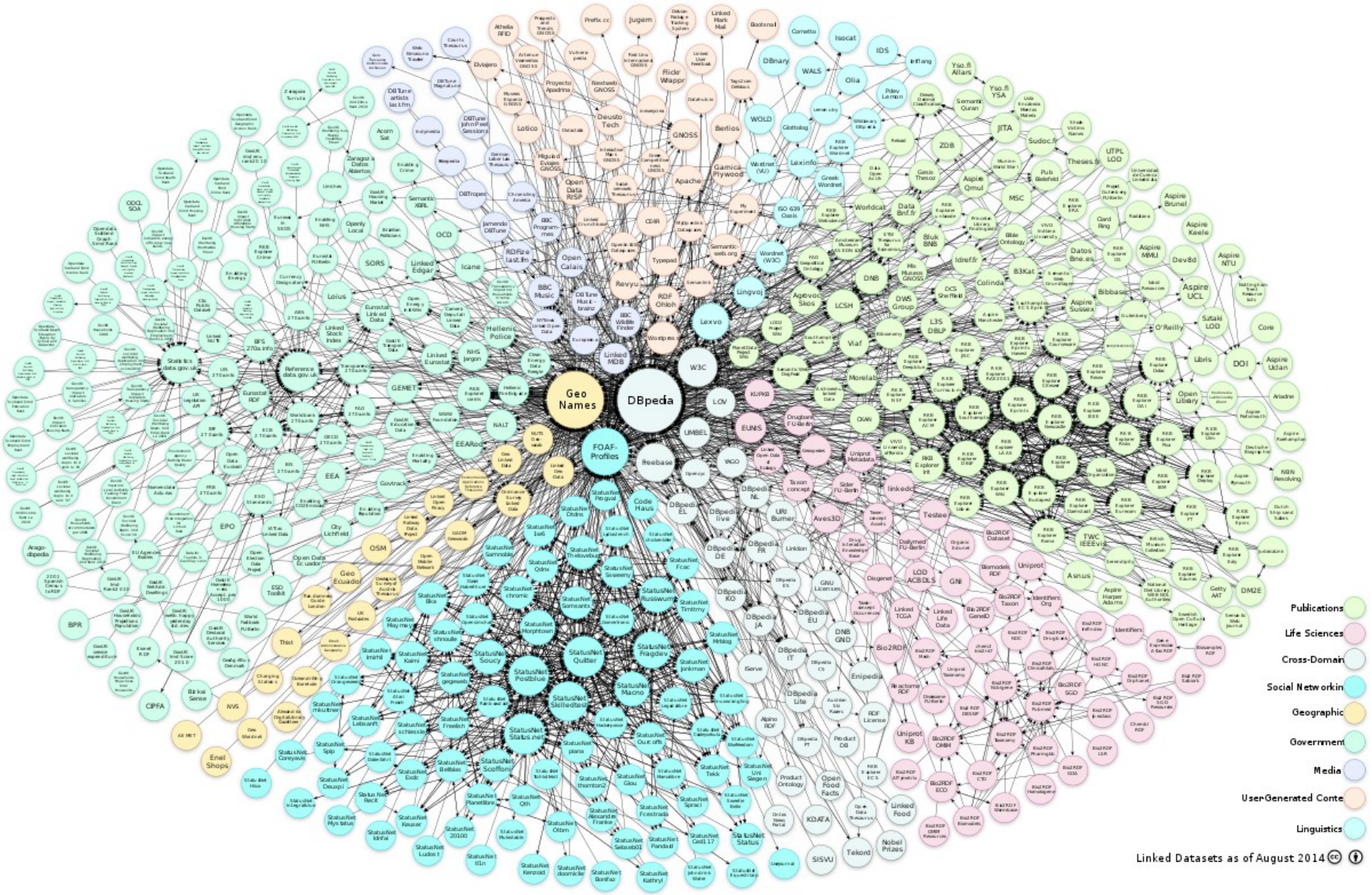
Retailer Vocabulary

"Mysterious" Vocabulary

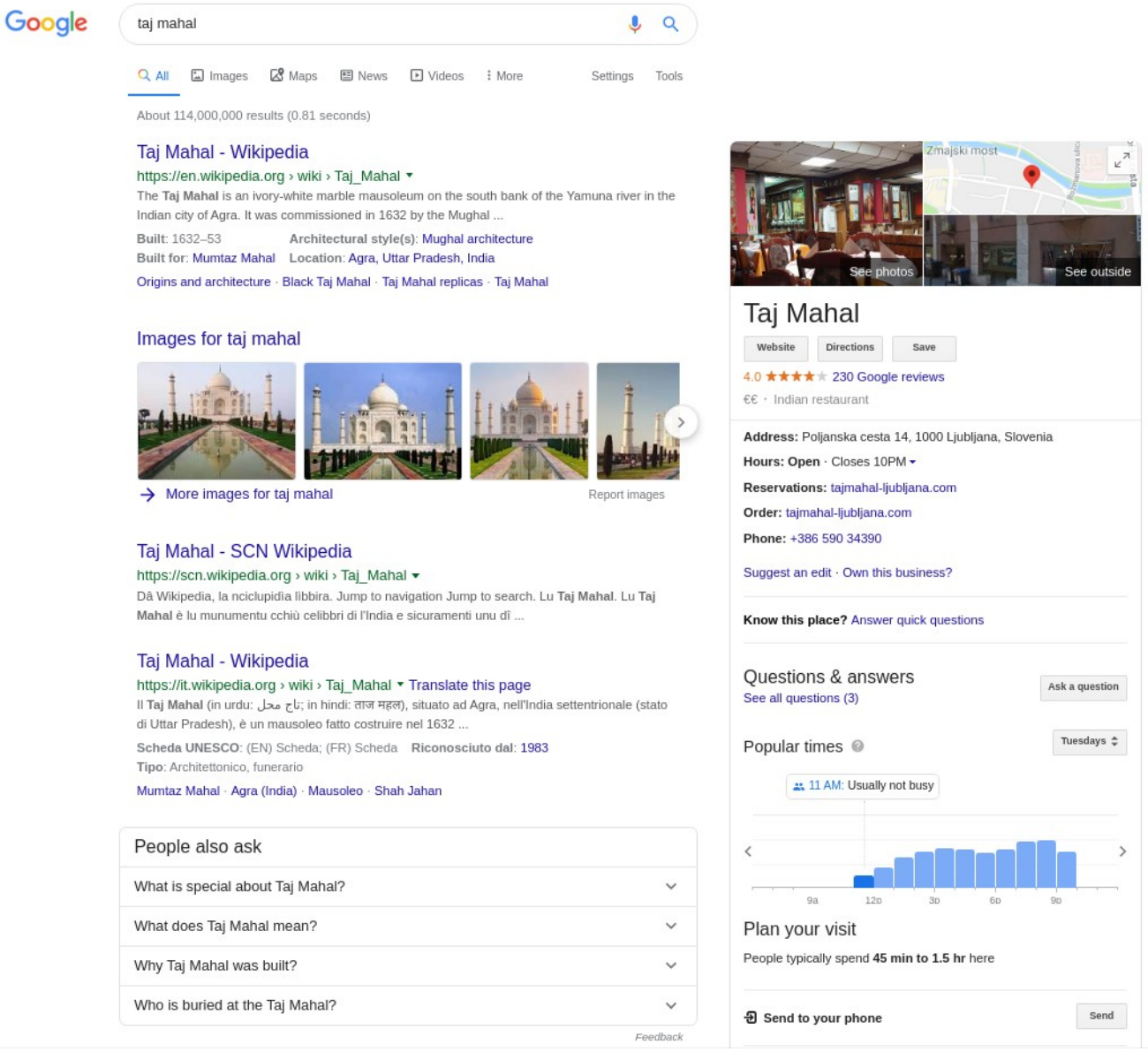
Example – DBpedia (1)

- The DBpedia project focuses on the task of converting Wikipedia content into structured knowledge
- Steps to build up DBpedia:
  - Convert Wikipedia content to RDF (Resource Description Framework) with metadata that describe the semantic context
  - Interlink DBpedia dataset with other open datasets
  - Develop interfaces and access modules

Example – DBpedia (2)



Example – Google's Knowledge Graph



Costs & Benefits of Open Data

- As a consumer:
  - You can access, look, print, store locally, share data
  - You can process and manipulate the data in any way you like
  - You can link to it from any other place
  - You can combine the data safely with other data
  - You can discover more (related) data while consuming the data
- As a publisher:
  - You might need converters or plug-ins to export the data from the proprietary format
  - You'll need to assign URIs to data items and think about how to represent the data
  - You need to either find existing patterns to reuse or create your own
  - You'll need to invest resources to link your data to other data on the Web
  - You may need to repair broken or incorrect links

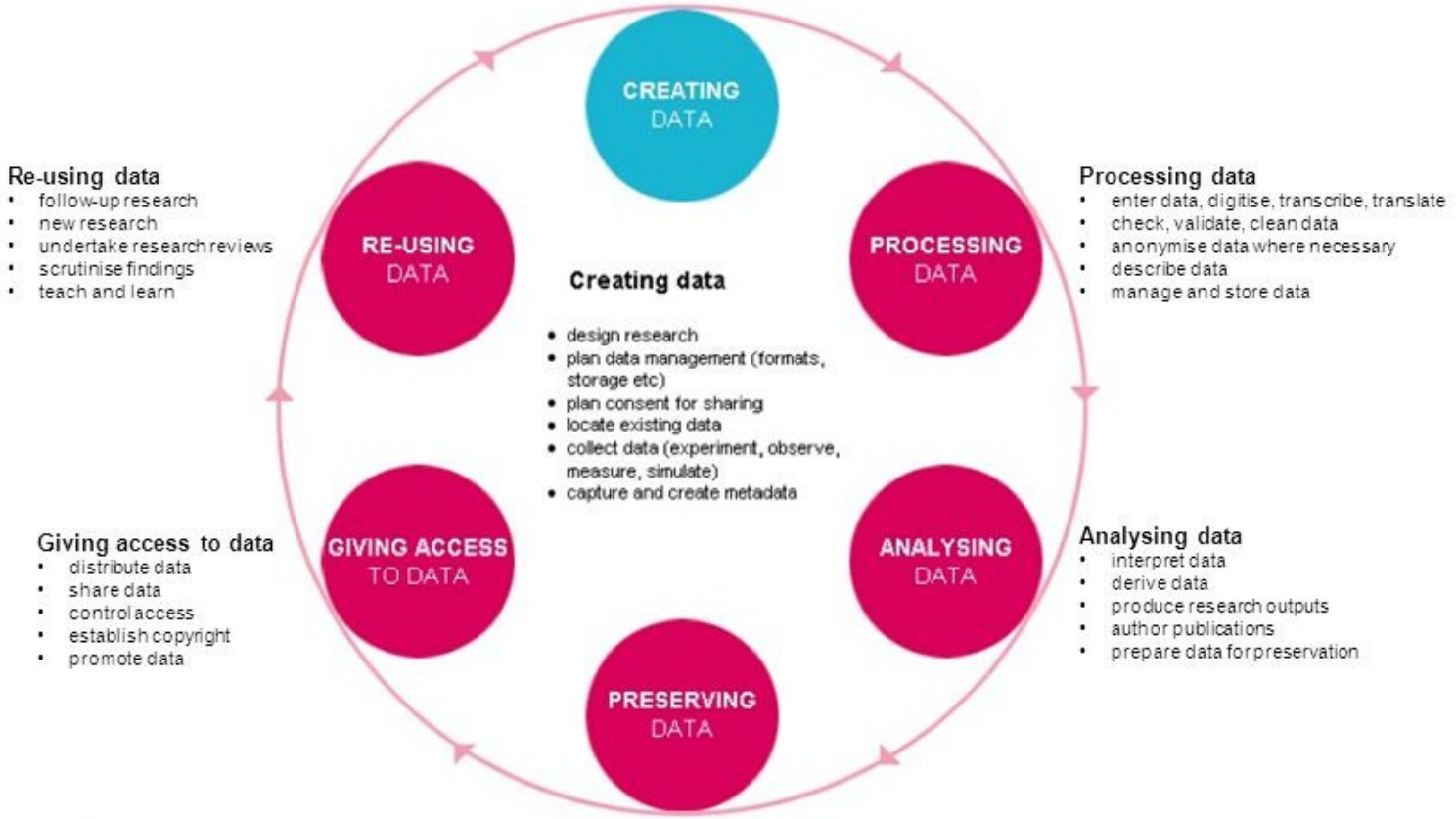
Metadata

- Metadata are data that describe other data
- For example *author*, *date created* and *date modified* and *file size* are very basic document metadata
- Metadata are data themselves
- Metadata are essential for:
  - Data description
  - Data discovery
  - Data linking
- Metadata must store all information necessary to understand and use data

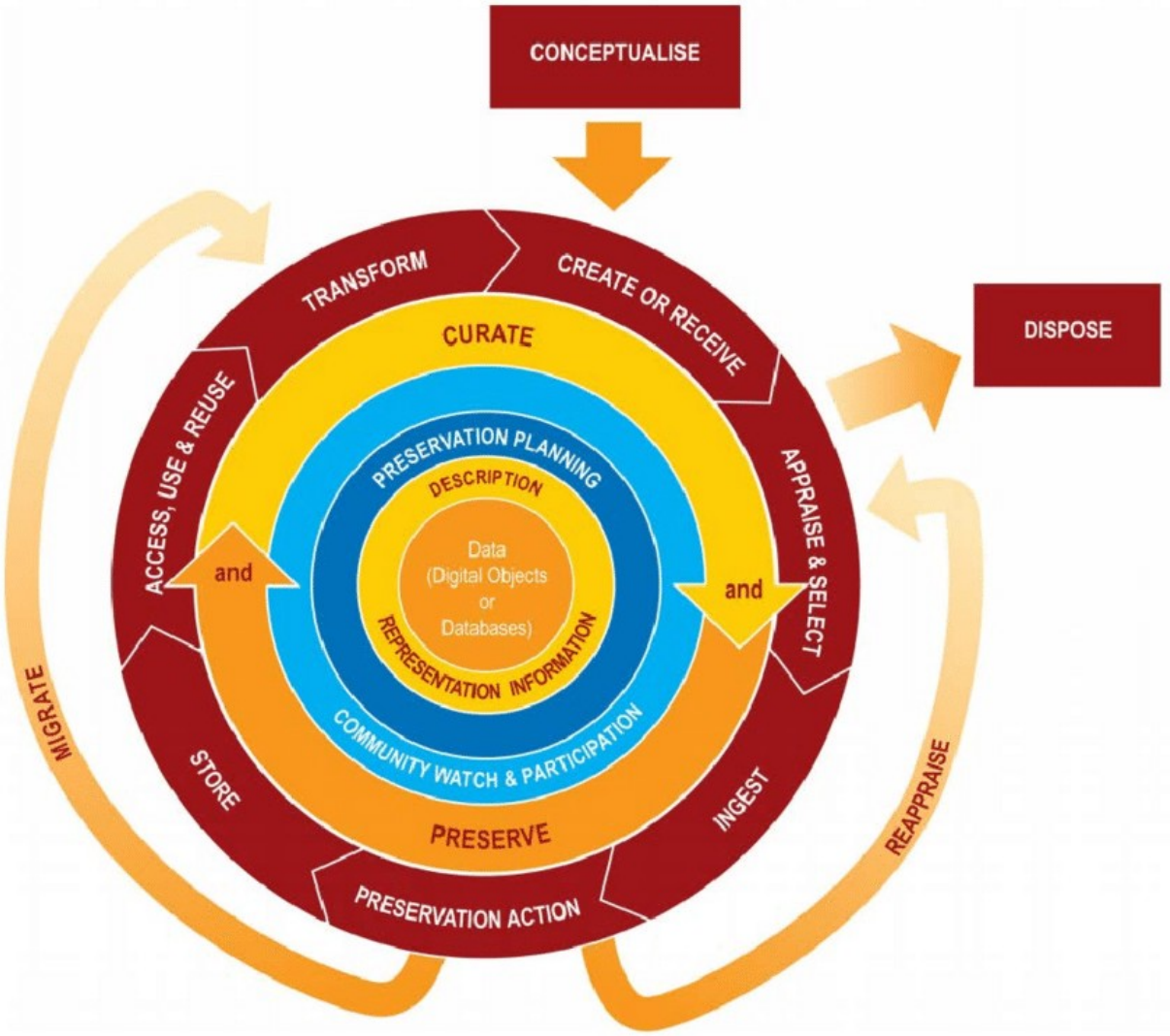
Data Management

- Providing Open Data requires a careful consideration of Data Management issues along the full Data Life Cycle
- Data Life Cycle involves several steps: creation, ingestion, archive, analysis, publishing, reuse
- Metadata Model are adopted according to different communities and infrastructure requirements
- There are different Data Life Cycle Models, such as:
  - UK Data Archive
  - Digital Curation Centre
  - US Geological Survey
  - ...and many others

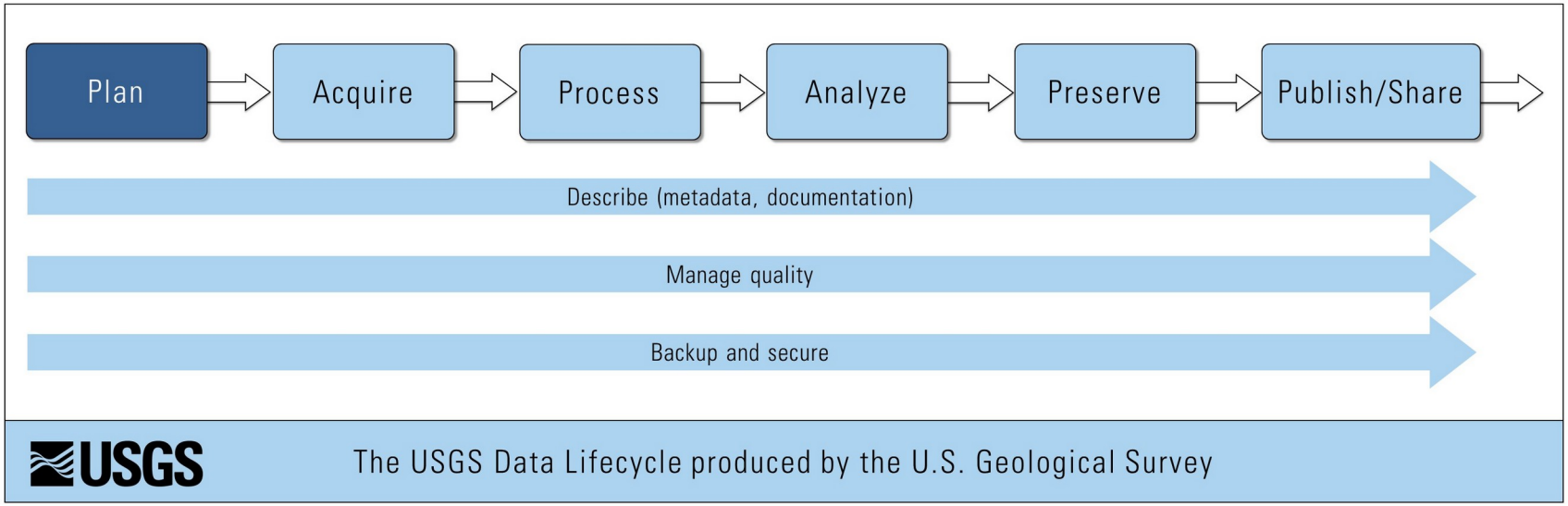
Data Life Cycle – UK Data Archive



Data Life Cycle – Digital Curation Centre



Data Life Cycle – US Geological Survey



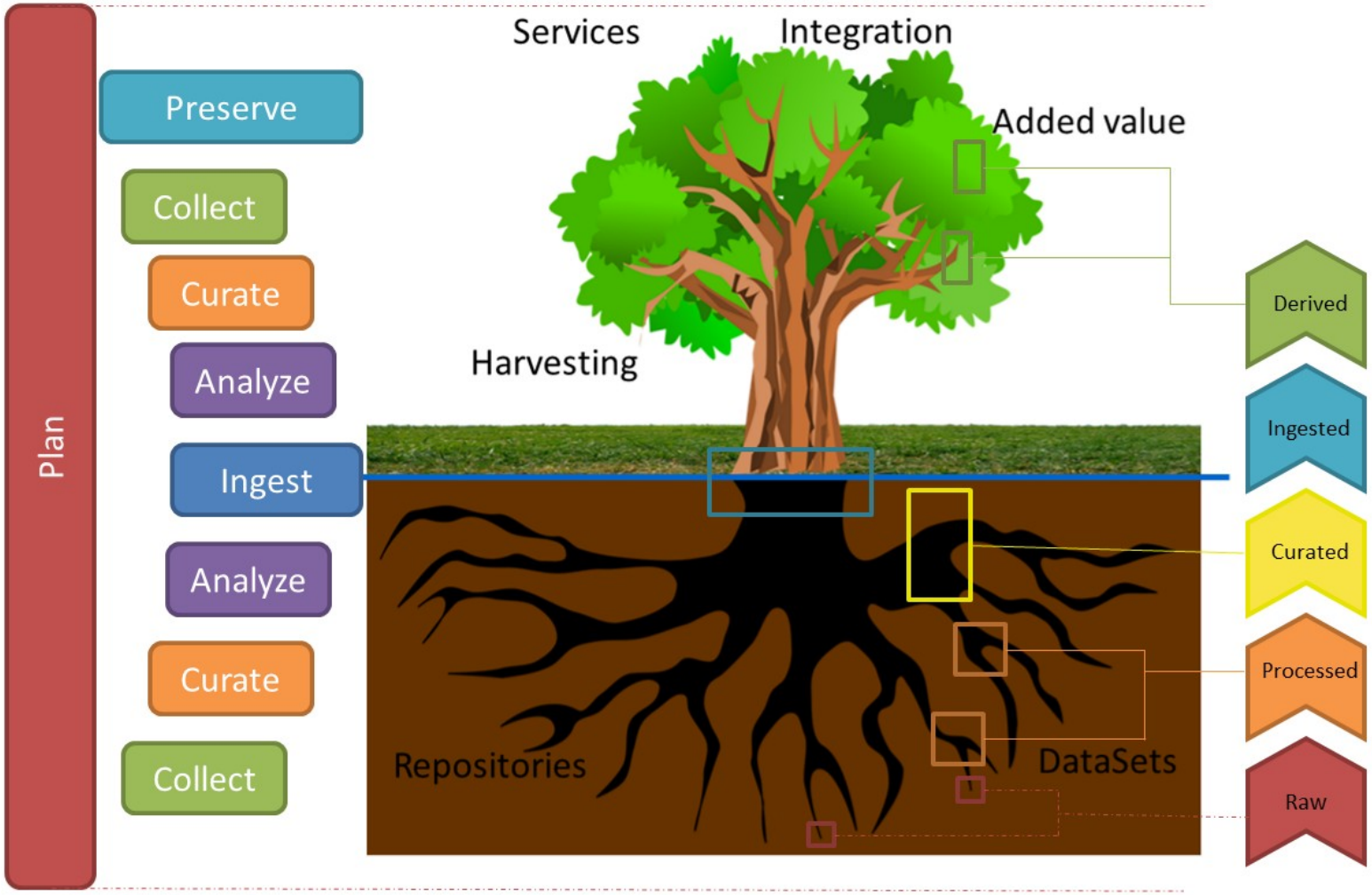
INDIGO-DataCloud – 6S

- INDIGO-DataCloud was an H2020 project aimed to develop an open source data and computing platform targeted at scientific communities, deployable on multiple hardware and provisioned over hybrid, private or public, e-infrastructures
- An unified view of the Data Life Cycle for all scientific communities is almost impossible given the high diversity in requirements, but INDIGO-DataCloud tried to define common practices to pave the way towards the European Open Science Cloud (EOSC)
- Six common stages (denominated 6S) in Data Life Cycle:
  - Stage 1 Plan: prepare a Data Management Plan (how gather data, metadata definition, preservation plan, etc.)
  - Stage 2 Collect: create and acquire raw data
  - Stage 3 Curate: perform actions on raw data to filter outliers, fix instrumental errors, and similar problems
  - Stage 4 Analyze: perform actions to give the data an added value and get new derived data
  - Stage 5 Ingest and Publish: associate data with metadata, assign persistent identifiers, publish in accessible repositories or catalogs
  - Stage 6 Preserve: store data, metadata and analysis for long-term

INDIGO-DataCloud – Data Levels

- Raw data: data taken by an instrument, sensor, human observation, etc. Instruments are usually considered to be calibrated before gathering data.
- Processed data: data is transformed in more useful units and some parameters (e.g. different sensors combination) are calculated.
- Curated data: data are filtered and all out-of-range data, human or instrument errors, outliers and other similar problems are corrected. Curation can be automatic or manual.
- Ingested data: datasets are prepared and transformed into a format suitable for distribution and re-use. A DOI (Digital Object Identifier) is assigned and proper metadata is associated to the dataset. The dataset can be published if desired, as it is also ready for external use
- Derived data: after applying an analysis method (model, simulation, statistical methods, etc.,) or integrating with other external or internal datasets, new derived data is generated, ready for publication, contributing to studies, or for further re-use. A new DOI and corresponding metadata may be assigned.

INDIGO-DataCloud – Arbor Metaphor



Flowers Metaphor

References (1)

- World Wide Web Consortium https://www.w3.org/
- Open Society Foundations https://www.opensocietyfoundations.org/
- Open Knowledge International https://okfn.org/
- Research Data Alliance https://www.rd-alliance.org/
- Open Science as a Practice http://openscienceasap.org/
- The Open Definition https://opendefinition.org/
- Budapest Open Access Initiative https://www.budapestopenaccessinitiative.org/
- 5 ★ Open Data https://5stardata.info/en/
- Tim Berners-Lee: The next Web of open, linked data https://www.youtube.com/watch?v=OM6XIICm_qo
- Tim Berners-Lee: Open, Linked Data for a Global Community https://www.youtube.com/watch?v=ga1aSJXCFe0
- What is Linked Data? https://www.youtube.com/watch?v=4x_xzT5eF5Q

References (2)

- Json http://www.json.org/ and http://json-ld.org
- RDF http://rdfa.info
- DBpedia https://wiki.dbpedia.org/
- Google's Knowledge Graph https://goo.gl/uu6bRM
- UK Data Archive http://www.data-archive.ac.uk
- DCC Lifecycle Model http://www.dcc.ac.uk/resources/curation-lifecycle-model
- U.S. Geological Survey https://www.usgs.gov/
- INDIGO-DataCloud https://www.indigo-datacloud.eu/
- INDIGO-DataCloud Deliverables D2.7 and D2.11 https://www.indigo-datacloud.eu/documents-deliverables