Keynote Presentation
National Open Science Plan for France:
From Strategy to Action
Paris, 4 December 2018
• Generalize open access to publications
• Structure research data and make it available through open access
• Be part of a sustainable European and international open science dynamic
Knowledge Infrastructures

“robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds”

Knowledge Infrastructures
Infrastructures

Opportunities in Open Science
New knowledge from old data

- https://library.cfa.harvard.edu/image-vocab/harvard-computers
Networks of data

http://humannaturelab.net/wp-content/uploads/2015/01/Fig1-no-text-village-2-only-selection.png
The Undiscovered: Many great discoveries in science are surprises.

https://www.radcliffe.harvard.edu/event/2018-undiscovered-symposium
Challenges in Open Science
Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship.

Center for Embedded Networked Sensing

- NSF Science & Tech Ctr, 2002-2012
- 5 universities, plus partners
- 300 members
- Computer science and engineering
- Science application areas

Slide by Jason Fisher, UC-Merced, Center for Embedded Networked Sensing (CENS)
Science <-> Data

Engineering researcher: “Temperature is temperature.”
Science <-> Data

Engineering researcher: “Temperature is temperature.”

Biologist: “There are hundreds of ways to measure temperature. ‘The temperature is 98’ is low-value compared to, ‘the temperature of the surface, measured by the infrared thermopile, model number XYZ, is 98.’ That means it is measuring a proxy for a temperature, rather than being in contact with a probe, and it is measuring from a distance. The accuracy is plus or minus .05 of a degree. I [also] want to know that it was taken outside versus inside a controlled environment, how long it had been in place, and the last time it was calibrated, which might tell me whether it has drifted.”

CENS Robotics team
LSST Timeline: https://www.lsst.org/about/timeline
Data Stewardship

Marie Curie's notebook aip.org

Pisa Griffin

Date:1/2/07.75 Place: Sakaltutan Zafar
He will grow old in his present house; new house is for sons - 5 sons. Not sure they want to live in village. He will only build another if they want him to. So came from Germany and did the plastering. He arranged the carpentry in Kayseri. Çok para gitti. (much money went) Has a tractor.

Date: July 1980 Place: Sakaltutan Zafar:
Household now Zafar and wife; Nazif Ural and wife and youngest son, still a boy. They run two dolmuş; one with a driver from Süleymanlı. Goes in and out once a day. He gets 8,000 a month. Zafar then said, keskin değil (not sharp - i.e. not profitable) I said he did very well on 8,000 TL with only two journeys a day. Nazif Ural has "bought" a Durak (dolmuş stop) from Belediye and works all day in Kayseri.
The DCC Curation Lifecycle Model provides a graphical high level overview of the stages required for successful curation and preservation of data from initial conceptualisation or receipt. The model can be used to plan activities within an organisation or consortium to ensure that all necessary stages are undertaken, each in the correct sequence. The model enables granular functionality to be mapped against it; to define roles and responsibilities, and build a framework of standards and technologies to implement. It can help with the process of identifying additional steps which may be required, or actions which are not required by certain situations or disciplines, and ensuring that processes and policies are adequately documented.

Data, any information in binary digital form, is at the centre of the Curation Lifecycle. This includes:

- **Simple Digital Objects** are discrete digital items; such as textual files, images or sound files, along with their related identifiers and metadata.
- **Complex Digital Objects** are discrete objects, made by combining a number of other digital objects, such as websites.

### Full Lifecycle Actions

**CONCEPTUALISE**
- **CREATE OR RECEIVE**
  - Conceptualise
  - Create or receive
  - Ingest
  - Appraise & select
  - Preserved
  - Transform
  - Store
  - Access, use & reuse

**PRESERVATION PLANNING**
- Description
- Representation information
- Preservation information
- Preservation planning

**PRESERVE**
- Preserve
- Preserve action
- Preserve store

**PRESERVATION ACTION**
- Preservation action
- Preservation store
- Preservation transform

**ACCESS, USE & REUSE**
- Access, use & reuse
- Retrieve
- Access store

**APPRaise & SELECT**
- Appraise & select
- Appraise & select store

**DISPOSE**
- Dispose
- Dispose store

**REAPPRAISE**
- Reappraise
- Reappraise store

**Migrate**
- Migrate
- Migrate store

**Store**
- Store
- Store store

**INGEST**
- Ingest
- Ingest store

**Transform**
- Transform
- Transform store

**Curate**
- Curate
- Curate store

**Community Watch & Participation**
- Community watch & participation
- Community watch & participation store

**Data** (Digital Objects or Databases)
- Data
- Data store

---

Assign administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long-term. Collect and assign representation information required to understand and render both the digital material and the associated metadata.

Plan for preservation throughout the curation lifecycle of digital material. This would include plans for management and administration of all curation lifecycle actions.

Maintain a watch on appropriate community activities, and participate in the development of shared standards, tools and suitable software.

Be aware of, and undertake management and administrative actions planned to promote curation and preservation throughout the curation lifecycle.

Conceive and plan the creation of data, including capture method and storage options.

Create data including administrative, descriptive, structural and technical metadata. Preservation metadata may also be added at the time of creation.

Receive data, in accordance with documented collecting policies, from data creators, other archives, repositories or data centres, and if required assign appropriate metadata.

Evaluate data and select for long-term curation and preservation. Adhere to documented guidance, policies or legal requirements.

Transfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements.

Undertake actions to ensure long-term preservation and retention of the authoritative nature of data. Preservation actions should ensure that data remains authentic, reliable and usable while maintaining its integrity. Actions include data cleaning, validation, assigning preservation metadata, assigning representation information and ensuring acceptable data structures or file formats.

Store the data in a secure manner adhering to relevant standards.

Ensure that data is accessible to both designated users and reusers, on a day-to-day basis. This may be in the form of publicly available published information. Robust access controls and authentication procedures may be applicable.

Create new data from the original, for example:
- By migration into a different format.
- By creating a subset, by selection or query, to create newly derived results, perhaps for publication.
Data Stewardship: the Reality

We just need to migrate the data from these systems to fit into that hole over there.

I’ll get the hammer.


Graduate students

Post-doctoral fellows
Data Stewardship: The Ideal

Pasquetto, I.V. (2018). *From Open Data to Knowledge Production: Biomedical Data Sharing and Unpredictable Data Reuses*. Phd Dissertation. [https://escholarship.org/uc/item/1sx7v77r](https://escholarship.org/uc/item/1sx7v77r)
"This is just the beginning": Using DNA and genealogy to crack years-old cold cases

Police are harnessing consumer DNA sites to solve old murders, which could spur a massive clearing of unsolved crimes.

by Kate Snow and Jon Schuppe / Jul.18.2018 / 4:30 AM ET

SHARE

POLICY FORUM | GENETICS AND PRIVACY

Genealogy databases and the future of criminal investigation

Natalie Ram¹, Christi J. Guerrini², Amy L. McGuire²
+ See all authors and affiliations

Science 08 Jun 2018:
Vol. 360, Issue 6393, pp. 1078-1079
DOI: 10.1126/science.aau1083

Paths Toward Open Science
Opportunities and Challenges

• Opportunities
  – Capture scientific products in digital form
  – Store, integrate, generate new knowledge

• Challenges
  – Skills and resources required to curate scientific records
  – Career paths for data science, curation work
  – Sustainability and stewardship of scientific products
  – Uses, reuses, and misuses of scientific products
Sustainable Open Science

• Create career paths
  – Data science
  – Curation and stewardship

• Commit to long-term infrastructure investments
  – Capture and sustain scholarly products
  – Stewardship of knowledge infrastructures

• Promote data reuse
  – Celebrate discovery
  – Anticipate controversy
  – Govern misuse