Data Archiving 101

Todd King

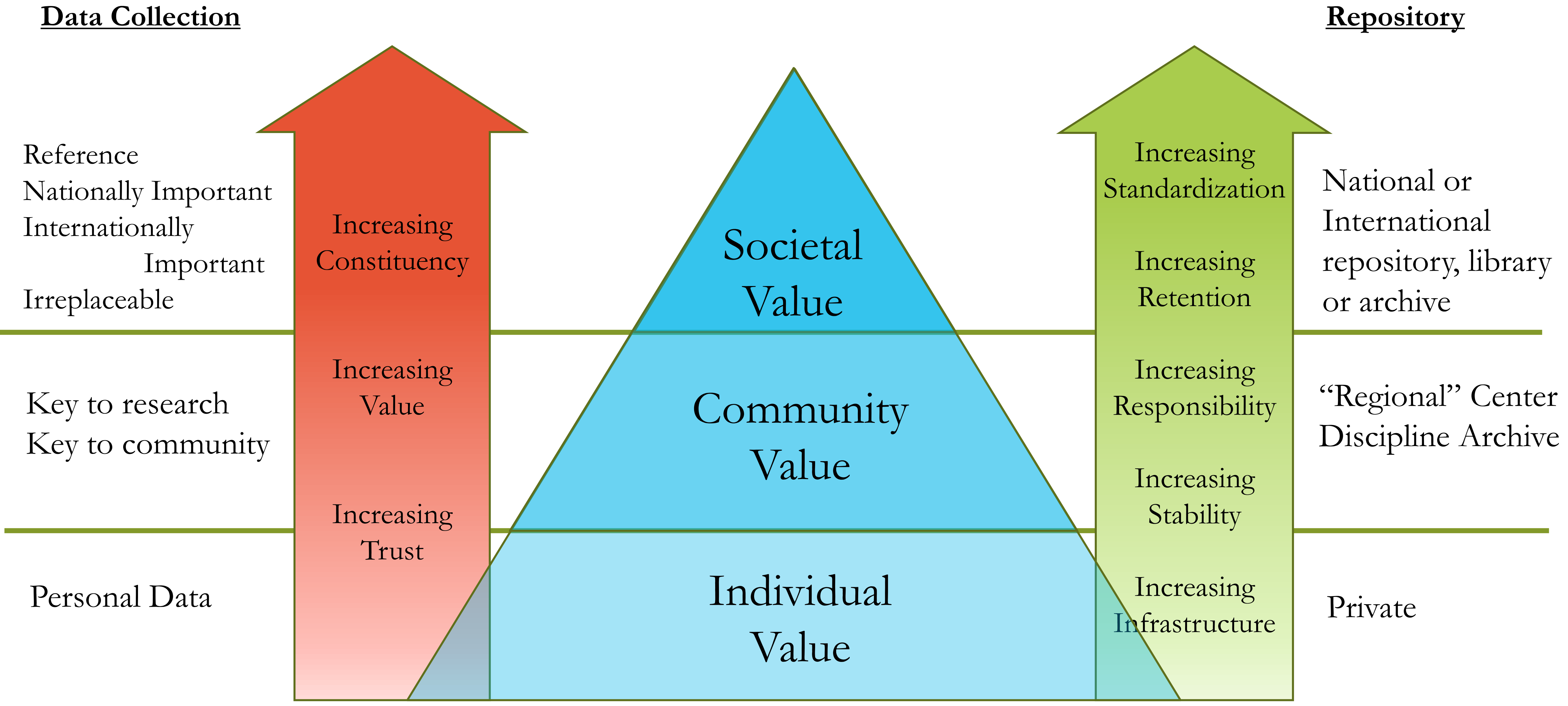
Data, Data Everywhere

The digital universe is **doubling in size every two years**, and by 2020 the digital universe – the data we create and copy annually – will reach 44 zettabytes, or 44 trillion gigabytes (IDC 2014 Report)

Much of the data is shared casually, but…

What to archive?

The Data Pyramid



Data Collection

Reference
Nationally Important
Internationally Important
Irreplaceable

Key to research
Key to community

Personal Data

Increasing Constituency

Increasing Value

Increasing Trust

Societal Value

Community Value

Individual Value

Increasing Standardization

Increasing Retention

Increasing Responsibility

Increasing Stability

Increasing Infrastructure

Repository

National or International repository, library or archive

"Regional" Center
Discipline Archive

Private

Based on figure by Francine Berman, Communications of the ACM, Dec 2008, Vol. 51, NO. 12, P. 53

As data value increases the responsibility to archive increases

The Government Viewpoint

- USA: Government Federal Records Act (44 U.S.C)
  - NASA: Full and open sharing of data
  - NSF: Share with other researchers … the primary data, samples, physical collections and other supporting materials
- EU: Directive 2013/37/EU (June 2013) on the re-use of public sector information.

Summary

If data is generated with public funds it should be preserved and shared as open data.

Archive Viewpoint

- **Durable Formats**
  - Open (patent free, unrestricted use)
  - Fully documented (free to share)
  - Minimal algorithms (structure over software)
- **Standardized Metadata**
- **Persistent Storage**
  - Self preserving (i.e. RAID 5)
  - Multiple copies (geographically separate)
- **Stewardship**
  - Quality assurance
  - Retention assurance

Curating Made Easy
or How to create an Archive

It’s more than saving the bits.

1. **Identify high value data**
   - Identify data (files) and other materials to be preserved
2. **Add metadata**
   - Assign a persistent identifier
   - Characterize (describe) the contents of a file with technical metadata
   - Include checksums
3. **Quality Assurance**
   - Review data and metadata for completeness and usability
4. **Place in a preserving repository**
   - Assure that it remains "complete and unaltered in all essential respects“ (fixity)
   - Retain versions
   - Protect against loss
5. **Access**
   - Provide open access
   - Trustworthy representations of what was originally received

Getting Started
What's Available

- **SPASE Metadata Model** (http://spase-group.org)
  - Supports Documents, Software, Numerical Data, Display Data, Simulations, Simulation Results
  - Supporting tools: editors, validators and generators.
  - International community of users.
- **HPDE Metadata Repository** (http://github.com/hpde)
  - Home of all NASA HPDE registered metadata (and some NOAA too)
  - Performs quality assurance checks before "publishing"
- **SPDF Data Repository** (https://spdf.gsfc.nasa.gov/)
  - Heliophysics mission data.
- **Any Other Data Repository**
  - SPASE can point to any web accessible repository

Open discussion

on how fun it is to archive.

References

- **White paper on NASA science data retention ,** https://nssdc.gsfc.nasa.gov/nssdc/data_retention.html
- **NASA Plan for Increasing Access to the Results of Scientific Research**, https://www.nasa.gov/sites/default/files/atoms/files/206985_2015_nasa_plan-for-web.pdf
- NASA Earth Science **Data & Information Policy,** https://science.nasa.gov/earth-science/earth-science-data/data-information-policy
- NSF **Dissemination and Sharing of Research Results**, https://www.nsf.gov/bfa/dias/policy/dmp.jsp
- **Open data: An engine for innovation, growth and transparent governance**, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0882:FIN:EN:PDF
- Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information, http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:02003L0098-20130717

Open Data

- 'Open data' describes data which is public, accessible at no or low cost and which can be reused or redistributed freely.