AUSTRALIAN NATIONAL DATA SERVICE

Provenance in Support of the ANDS Four Transformations

Mingfang Wu Andrew Treloar

Outline

- ANDS overview
- Why Provenance
- Provenance Activities
- Lessons Learned
- Future Plans

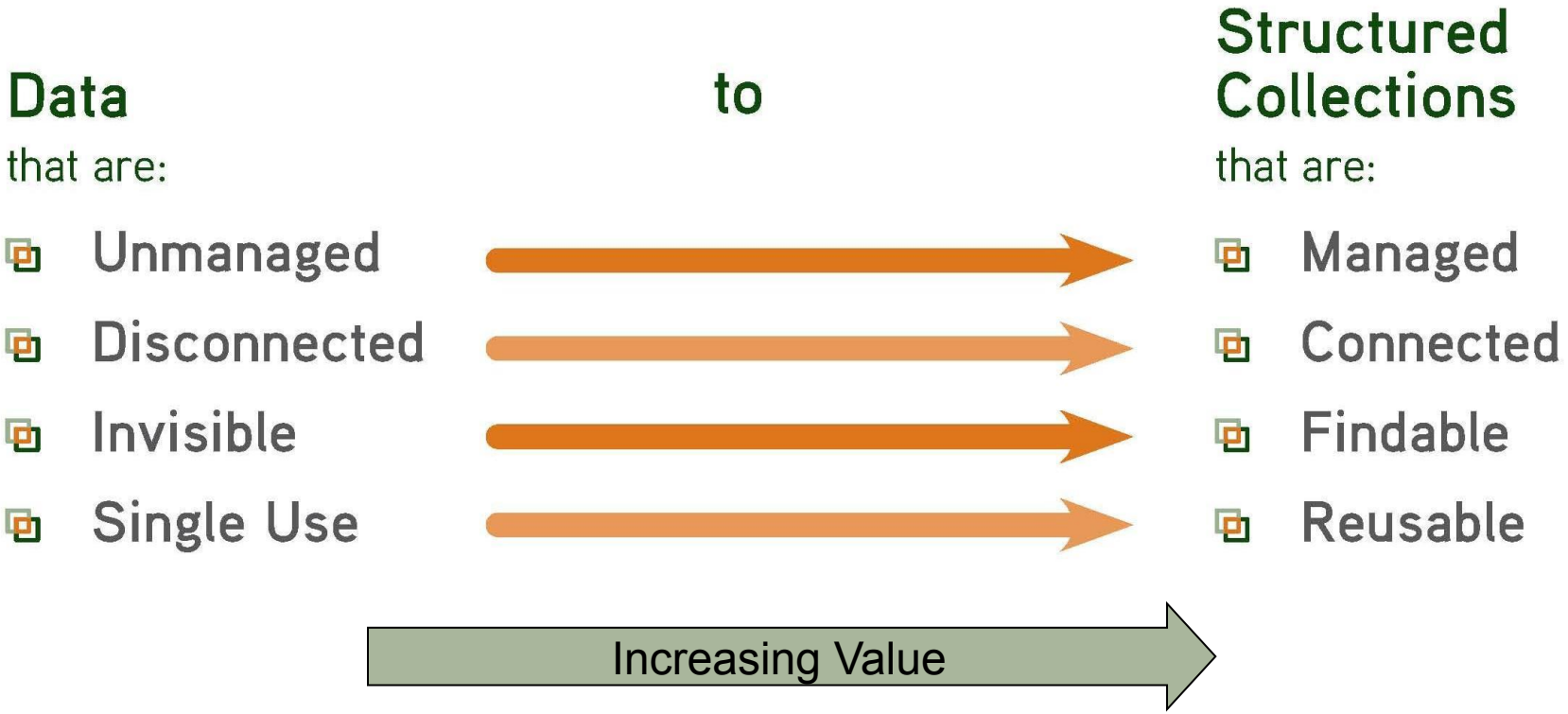
ANDS at a glance

- In operation since 2009
- Approximately AUD90M total investment
- 45 staff (mostly Melbourne, Canberra)
- Working to make Australia's research data more valuable
- Funding through to mid 2016 (probably 2017)

Key differentiators for ANDS

- No actual data storage
- Nationally co-ordinated approach
- Institutionally-focussed engagement
  - "helping them meet their research data ambitions"
- Engaging with large nationally-funded discipline investments
- Bulk of funds spent outside ANDS
  - https://projects.ands.org.au/getAllProjects.php?start=all
- All disciplines covered
- Focus on adding value to data and re-use

How Do We Make Data More Valuable?



So that researchers can easily publish, discover, access and use research data through the Australian Research Data Commons

Four Transformations and Provenance

- Unmanaged to Managed
  - Capture provenance as early as possible
- Disconnected to Connected
  - Provenance provides many possible connection points
- Invisible to Findable
  - Provenance provides possible alternative discovery paths
- Single-Use to Reusable
  - Provenance can be useful for re-analysis

Provenance-related activities

- Funded projects
  - provenance service and/or provenance integration
- Mapping RIF-CS to PROV-O
- Linking provenance information to data collection descriptions
- Local community coordination
- Engagement with Research Data Alliance IG

Lessons Learned

- Good provenance capture practice is good data management practice
- Installing a provenance management system is the easy bit
  - adopting persistent identifiers for all of the components to which the provenance will refer is much tougher
- Provenance may require adjustments to researcher behaviour
  - “I don’t want to expose my secret recipe!”

Future activity

- Working with discipline partners on domain-specific PROV-O extensions
- Tighter connections between Australian Research Data Provenance group and RDA IG
- Provenance – not just for machines!

Provenance Information

Data creation date: September 26th, 2011
Data creator(s): Gad Abraham
Data is derived from: Five human breast cancer microarray gene expression datasets, Five human gene sets from MSigDB Molecular Signatures Database
Data is produced by: Computational Model for Gene Set Analysis to predict breast cancer prognosis based on microarray gene expression data
Lineage statement: (if there is any ...)
More provenance information: http://data.nicta.org.au/service/provenance/search?entity=https://nicta.org.au/prov/vis-id=1234
Note: This provenance query will return a visulisation of provenance map centralised on this data collection

Related Publications

Abraham et al, 2010 "Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context" BMC Bioinformatics 11:277.
doi : 10.1186/1471-2105-11-277

The publication is supported by the dataset

Related Websites

Creative Commons Attribution 3.0 Australia License
URI : http://creativecommons.org/licenses/by/3.0/au/

Identifiers

Local : www.nicta.com.au/collection-3
DOI : doi:10.4225/02/4E9F69F7AE206

Questions?

- andrew.treloar@ands.org.au
- mingfang.wu@ands.org.au

- andrew.treloar.net