The Long Tail of Data Wagging the Institutional Repository

Open Repositories 2013

Chuck Humphrey
University of Alberta

Research data

- Is everything that is digital also data?
- There is digital content that has research potential but is not research data.
- Research data are the products that provide evidence in the research process.

Various stages of development

- We are all at various stages of dealing with research data and our repositories
  - Some may not yet deal with research data in their repository but are now investigating how to go about incorporating research data into their digital collections
  - Some may have started ingesting research data and are now looking at their next steps in this area
  - Some may have well established research data collections and are looking at ways to collaborate with other repositories or at how to fit into the emerging global research data ecosystem

The challenges of research data

- The heterogeneous nature of research data brings challenges to repositories in the following areas:
  - Policy foundation
  - Extent of processing for ingest and the resulting workflow
  - Formats
  - The design of the Archival Information Package
  - Identity of the repository
  - Skilled professionals

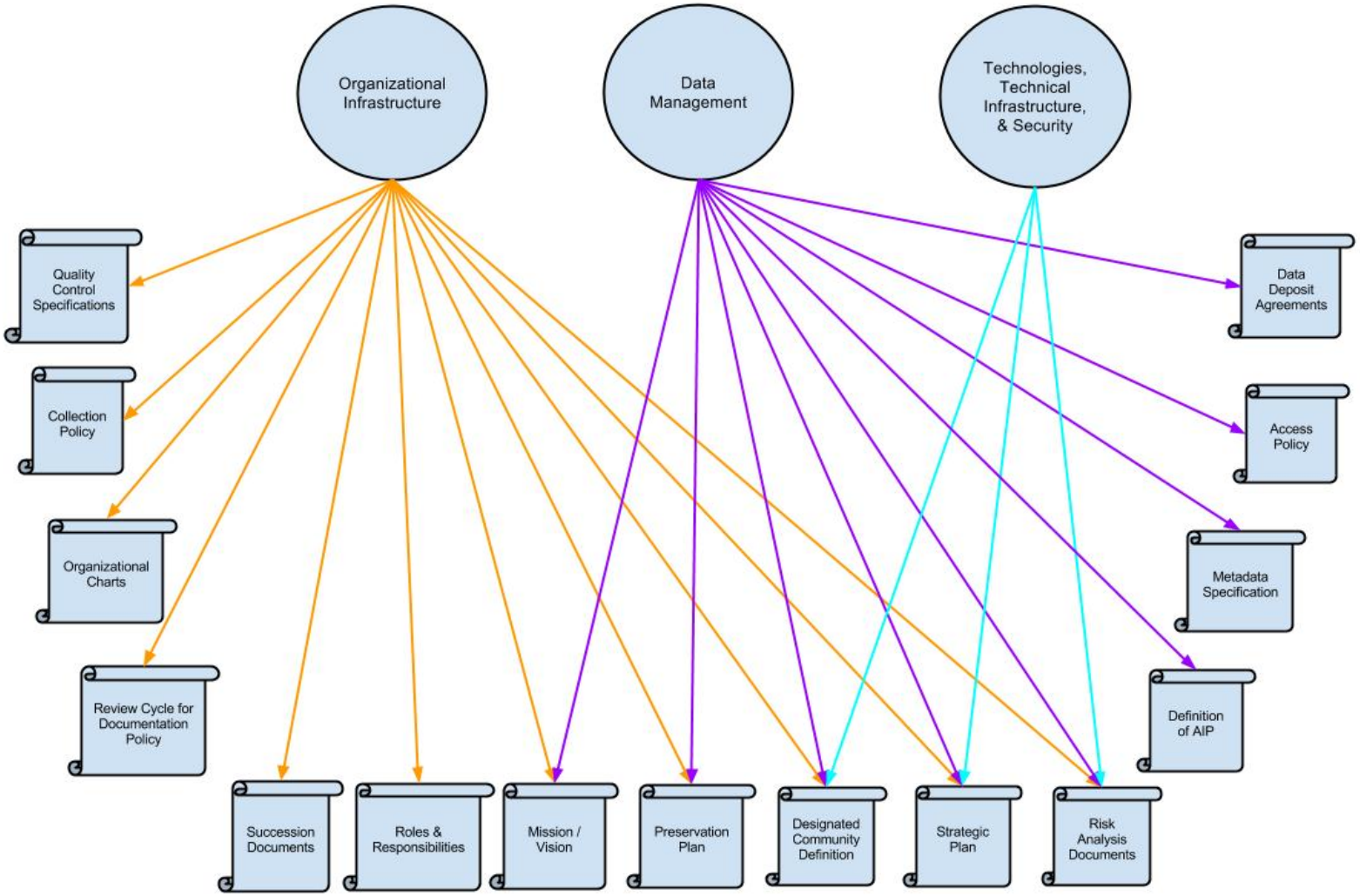
Two major environmental drivers

- Two significant sources of motivation for increased organizational interest in better managing research data
  - e-Science movement, now generalized to e-Research and expanded by Jim Gray's Fourth Paradigm argument
  - Academic integrity and interests in the replication of research findings
- These drivers carry different expectations
  - Collection versus product
  - Interoperable versus reproducible

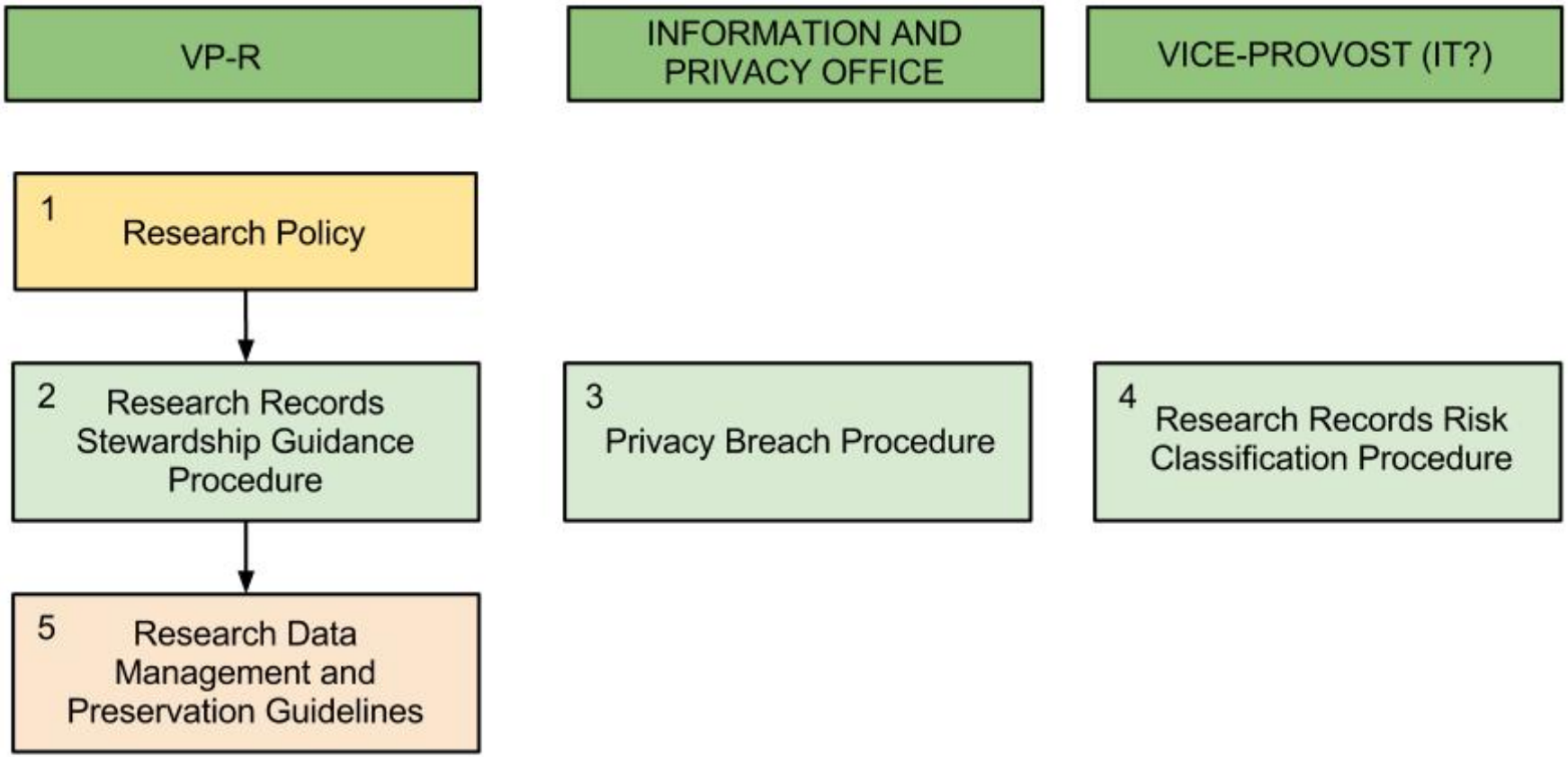
Data and a policy foundation

- Assumption: a set of policies exists based on the TRAC checklist or on an adaptation of the OAIS Reference Model
- Policies will likely need to be modified to support research data, especially if just getting into data
  - This is illustrated in the next slide, which shows the policy documents framework from the report by the Canadian Polar Data Network to the Canadian High Arctic Research Station on scientific and technological research data management infrastructure

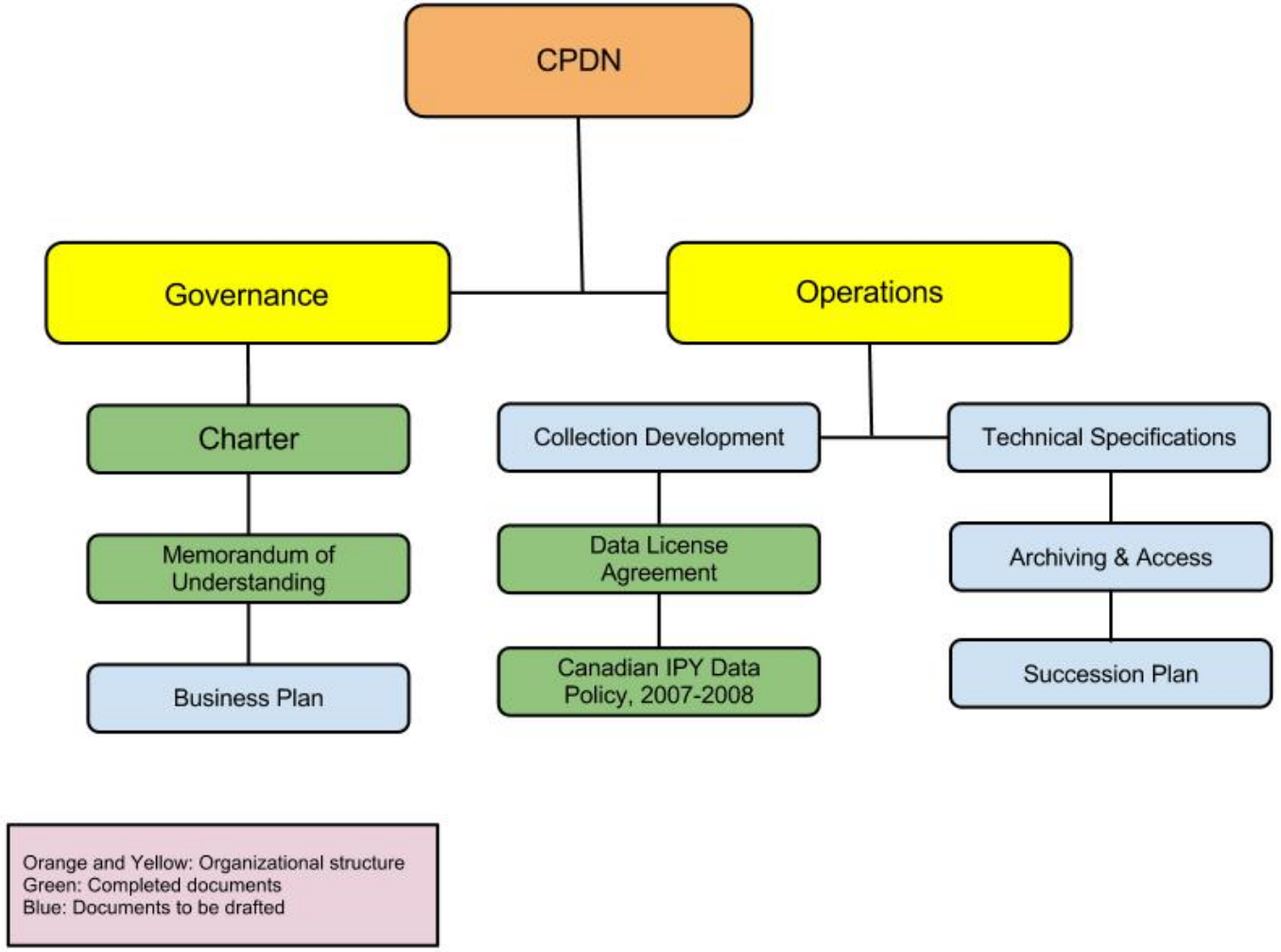
Data policy document framework



Data policy, procedures & guidelines



Data policy document framework



Extent of processing prior to ingest

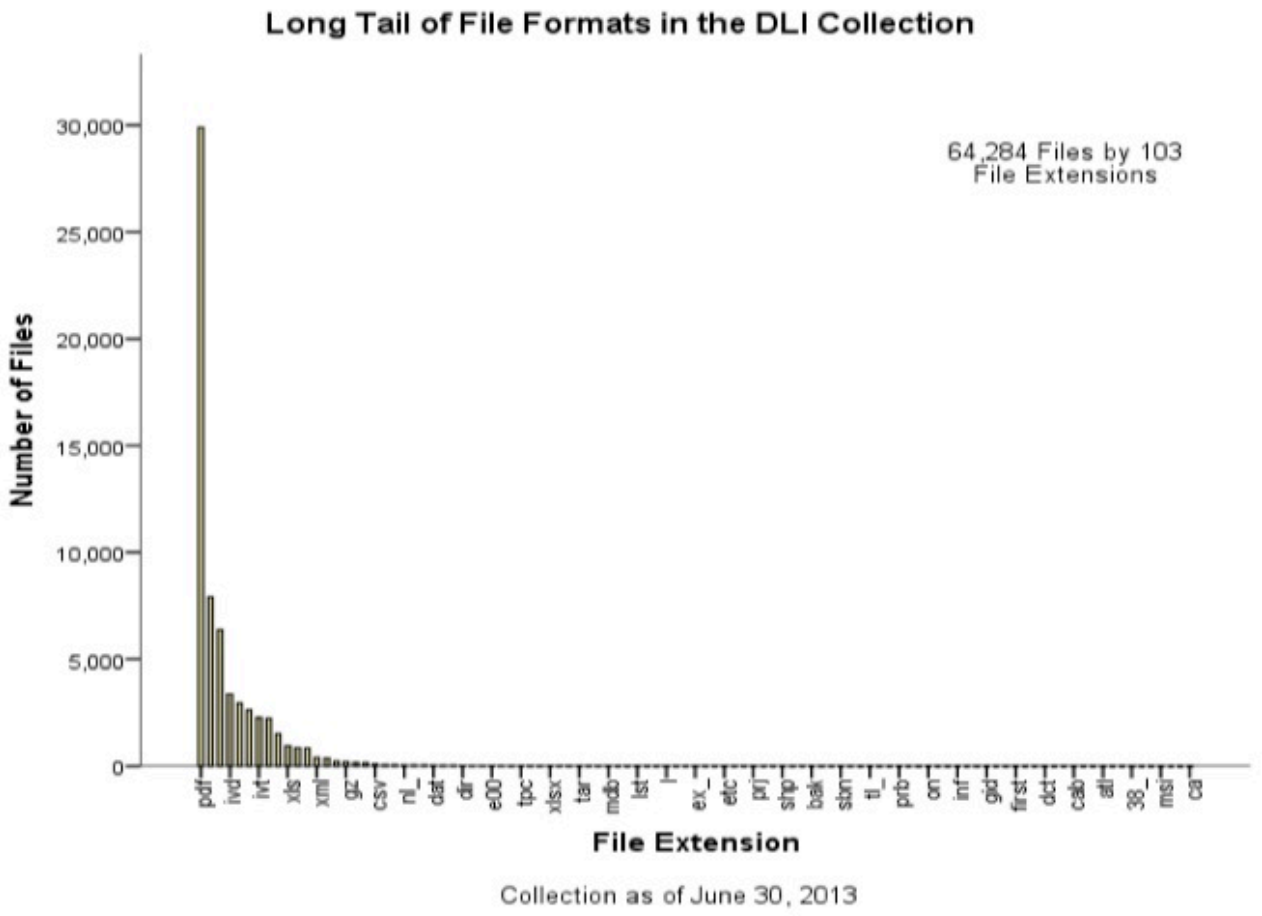
- Best practices for preparing data for ingest exist through some well-established domain archives, such as the ICPSR and UK Data Archive
  - Policy should guide whether research data get additional processing prior to ingest
    - Decision to accept "as is" or to do additional processing
  - Processing are steps needed to ensure completeness of documentation and data files, to screen for sensitive information, to conduct quality evaluations, to assign administrative content, to prepare generalized formats, etc.
  - Obstacles
    - Not being able to provide guidance soon enough in lifecycle
    - Not having an adequate data curation toolkit

Research data formats

- This is complicated by the wide range of analytic software formats that researchers and data producers use and by the unusual or inappropriate naming conventions that are employed
- The next slide shows the long tail of file formats in the repository for the Statistics Canada Data Liberation Initiative (DLI)
  - Files are received "as is" from STC author divisions, except for some of the production of SPSS syntax files to read microdata files

Research data formats: an example

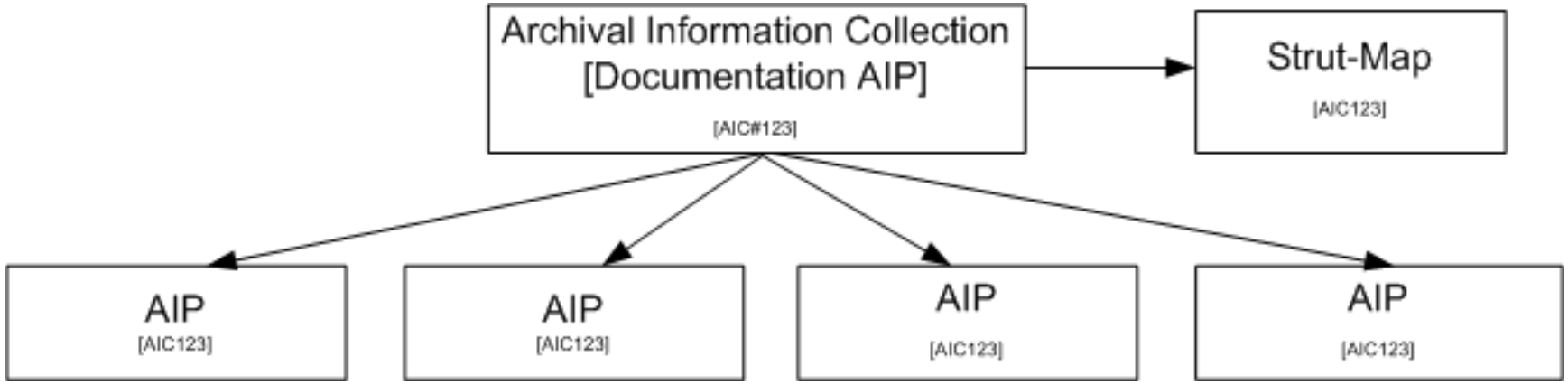
	extentype	N_FILES	Percent
2	pdf	29899	46.51
3	zip	7922	12.32
4	exe	6381	9.93
5	ivd	3363	5.23
6	txt	2948	4.59
7	doc	2620	4.08
8	ivt	2276	3.54
9	sps	2239	3.48
10	sas	1508	2.35
11	xls	945	1.47
12	tgz	851	1.32
13	wpd	839	1.31
14	xml	388	0.60
15	ppt	350	0.54
16	htm	222	0.35
17	gz	198	0.31
18	rtf	165	0.26
19	wp	148	0.23
20	csv	93	0.14
21	can	58	0.09
22	dl_	56	0.09
23	nl_	51	0.08
24	dll	50	0.08
25	gif	46	0.07
26	dat	42	0.07
27	dbf	41	0.06
28	jpg	34	0.05
29	dir	31	0.05
30	dwt	28	0.04
31	tpl	26	0.04
32	e00	26	0.04
33	cat	23	0.04
34	fra	22	0.03
35	tpc	18	0.03
36	spc	18	0.03
37	hlp	18	0.03
38	xlsx	17	0.03
39	png	16	0.02



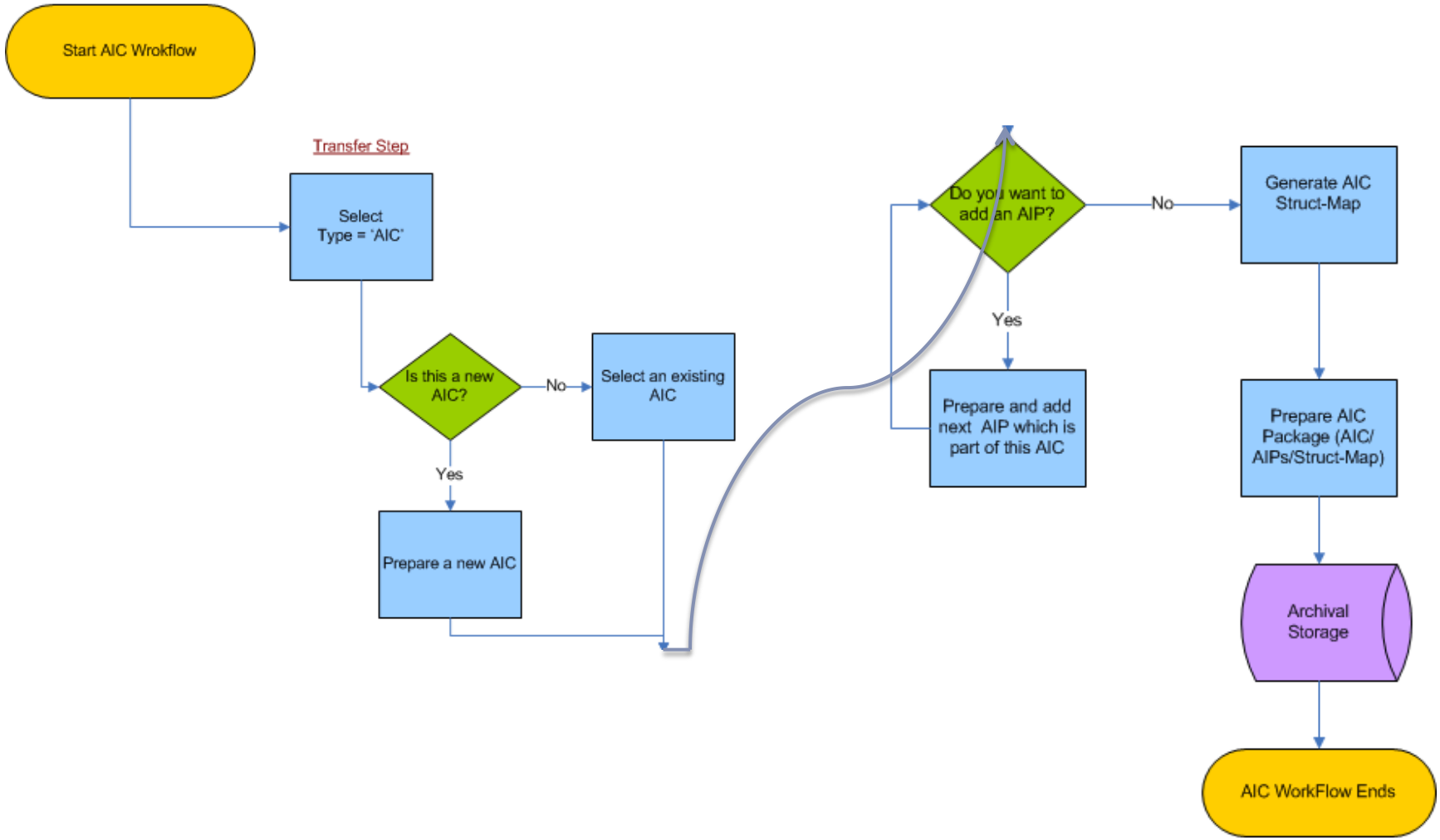
The AIP for research data

- Research data require thought as to the design of the Archival Information Package (AIP), that is, to the digital object produced from the Submission Information Package (SIP) and that is placed in archival storage
- Because the context in which research data are produced is vitally important for others to understand the data, efforts are made to document as much context as possible
  - Integrating contextual information with the research data needs to be considered in the AIP design
  - Multiple data files that are related also need thought

AIP design for research data: example



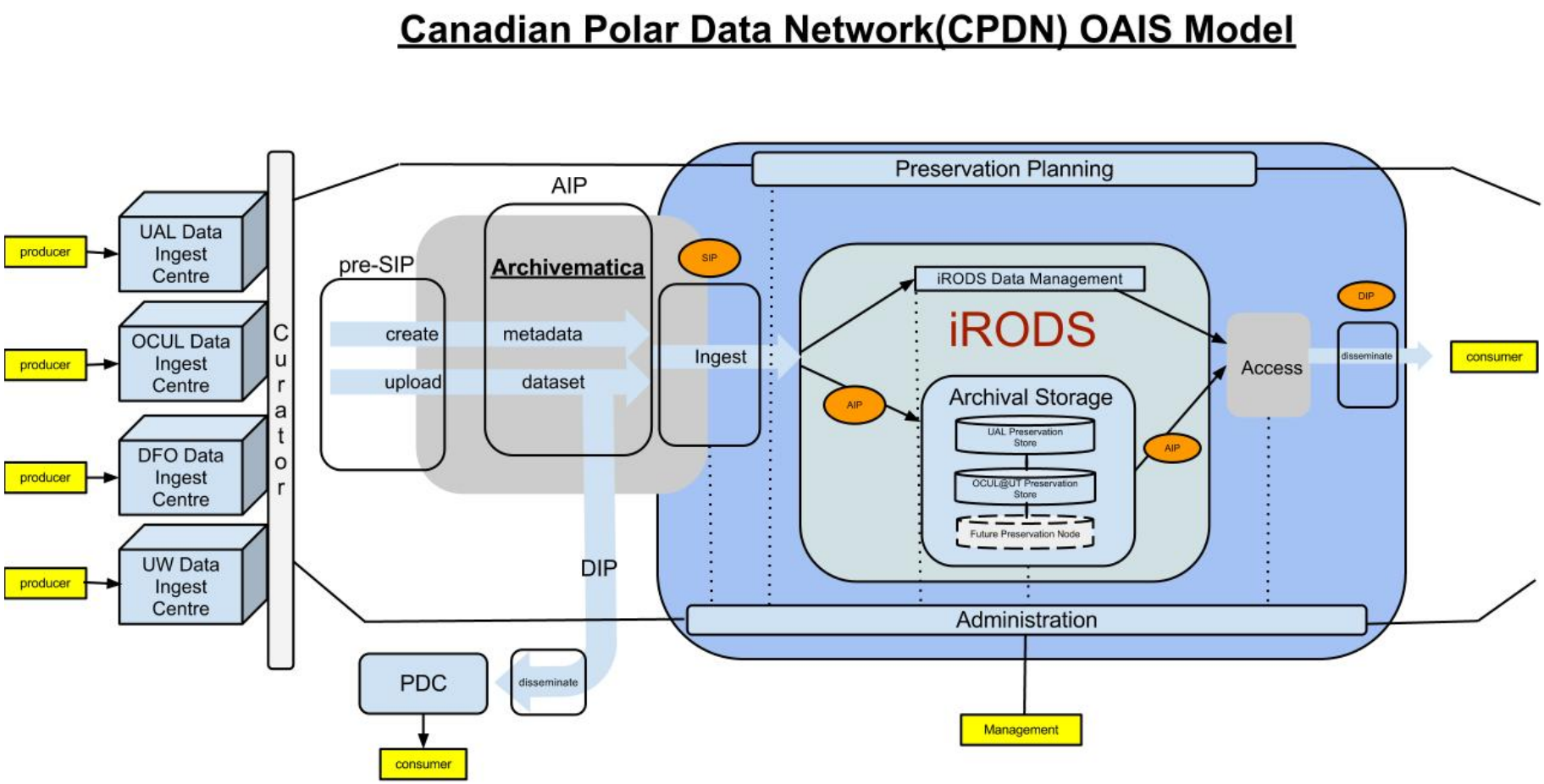
AIP data workflow: example



Repository: as a brand

- Do you think of your repository as a digital collection or as a platform?
- Quaecumque Vera : whatsoever things are true
- ERA : Educational and Research Archive
  - ERA-data
  - ERA-theses
  - ERA-text
- Collection distinctions should help direct decisions around repository infrastructure and services
- The next slide is an example of a mixed infrastructure model to support the data repository for the Canadian Polar Data Network

Mixed infrastructure model for data



Research data curation expertise

- Build a team environment for data curation
  - While not a sustainable solution, strategically select research projects to serve as an embedded data curator
  - Within the Library, develop a team of experts
    - U of Alberta example: Digital initiatives coordinator, Preservation officer, Digital initiatives technology librarian, Digital initiatives applications librarian, Institutional repository librarian, Metadata librarians, GIS librarian, Data library coordinator, Data curator intern
    - Develop liaison librarian roles (mainstreaming research data)
    - Capitalize on Co-op, Intern, and Post-doc data curators

Summary

- The special requirements of research data need to be rooted a repository's policy foundation
- Preparing research data for submission often requires additional processing, degrees of intervention, mediation, best practices in data management, policy support, and a data curation toolkit
- The variety of formats for research data requires a community effort to manage
- The design of the AIP is important in creating sound digital objects for research data (don't defer to technology on this point!)
- Research data should have its own identity in a repository of mixed digital content
- Build data curation teams with complementary expertise