Lecture 23 – Data Curation and Preservation

Open Data Management & the Cloud

(Data Science & Scientific Computing / UniTS – DMG)

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

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

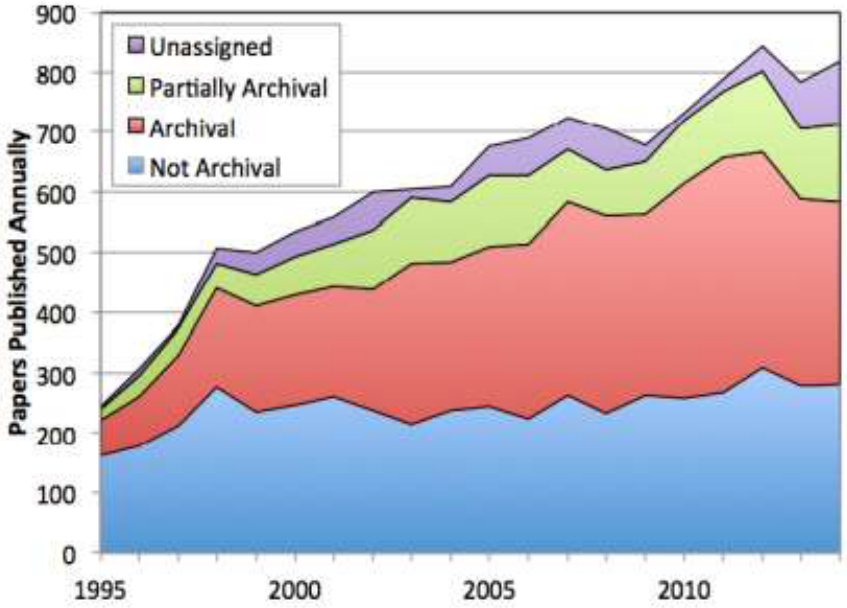
Open Data and Access in Science

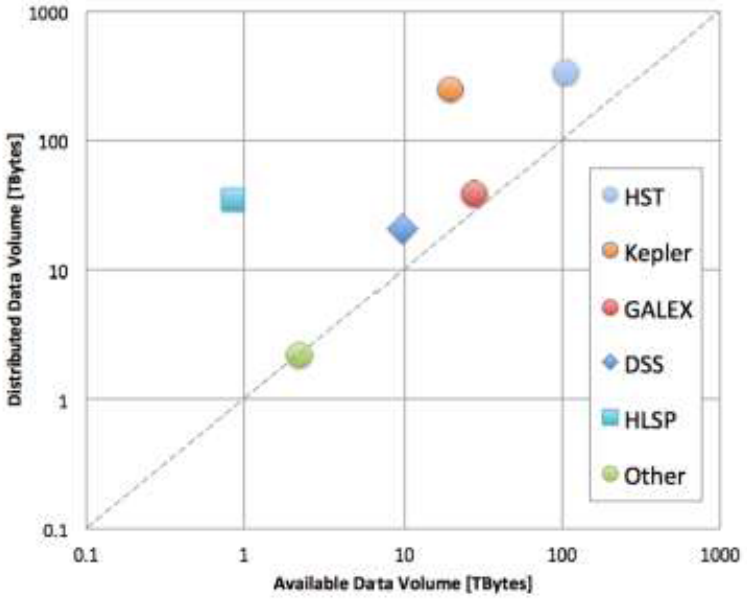
- Once published, scientific data should remain available forever so that other scientists can reproduce the results and do new science with the data.
- Data may be used long after the project that gathered it ends.
- Later users will not implicitly know the details of how the data was gathered and prepared.
- To understand the data, those later users need the metadata:
  - how the instruments were designed and built;
  - when, where, and how the data was gathered;
  - a careful description of the processing steps that led to the derived data products that are typically used for scientific data analysis.
- It is now feasible, even economical to store everything from most experiments.
  - But, documenting and curating the data is certainly not free.

Why preserve digital data?

- Digital data preservation should be a key aspect of all research projects
  - only by referring to verifiable data can your research be judged as sound
  - some research data are unique and cannot be replaced if destroyed or lost
- A tremendous growth in computational power, and in networking bandwidth and connectivity, has resulted in an explosion in the number of organizations making digital information available.
- Preserving information in digital forms is much more difficult than preserving information in forms such as paper and film
  - Digital data advantages:
    - searchability
    - replication
  - Digital data disadvantages:
    - rapid obsolescence of digital technologies
    - risk of losing the possibility of restoring, rendering or interpreting the information
    - organizational, legal, industrial, scientific and cultural issues
- In some case we need to provide a framework that may be expanded by other efforts to cover Long Term Preservation of information that is NOT in digital form (e.g., physical media and physical samples)

Why archiving is important?





- 60% of MAST papers based, in whole or in part, on archival (HST) data

What data should be preserved?

- Some data are irreplaceable and must be saved (**ephemeral**); other data can be regenerated (**stable**)
- Ephemeral data must be preserved, cannot be reproduced or reconstructed
  - If no one records them today, in a decade no one will know today's rainfall, sunspots, ozone density, or oil price.
- Stable data does not need to be saved, but there is an economic tradeoff between preserving it or recomputing/remeasuring stable data.
  - Data derived from simulations, from reductions of other data, or from measurements of time-invariant phenomena.
- The metadata about derived data products are ephemeral:
  - The design documents, email, programs, and procedures that produce a derived dataset would all be impossible to reconstruct.
- Given the metadata, the derived data can be reconstructed from the source data, so derived data are stable. One need only record the data reduction procedures in order to allow others to reconstruct the data.
- In summary, ephemeral data must be preserved; stable data need not be preserved. Metadata is ephemeral.

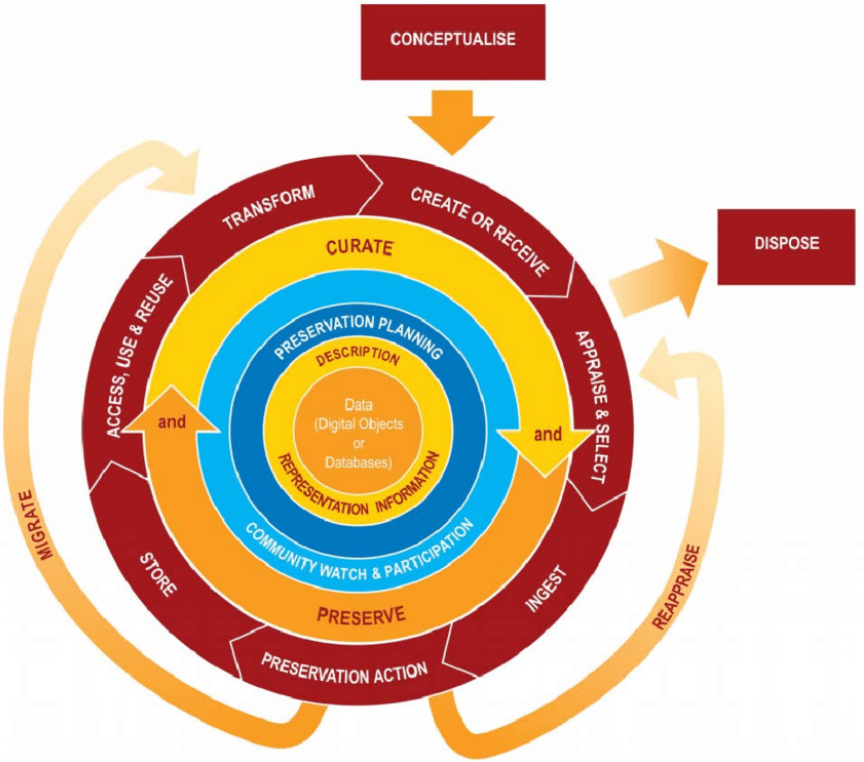
Metadata

- 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

What are digital curation and preservation?

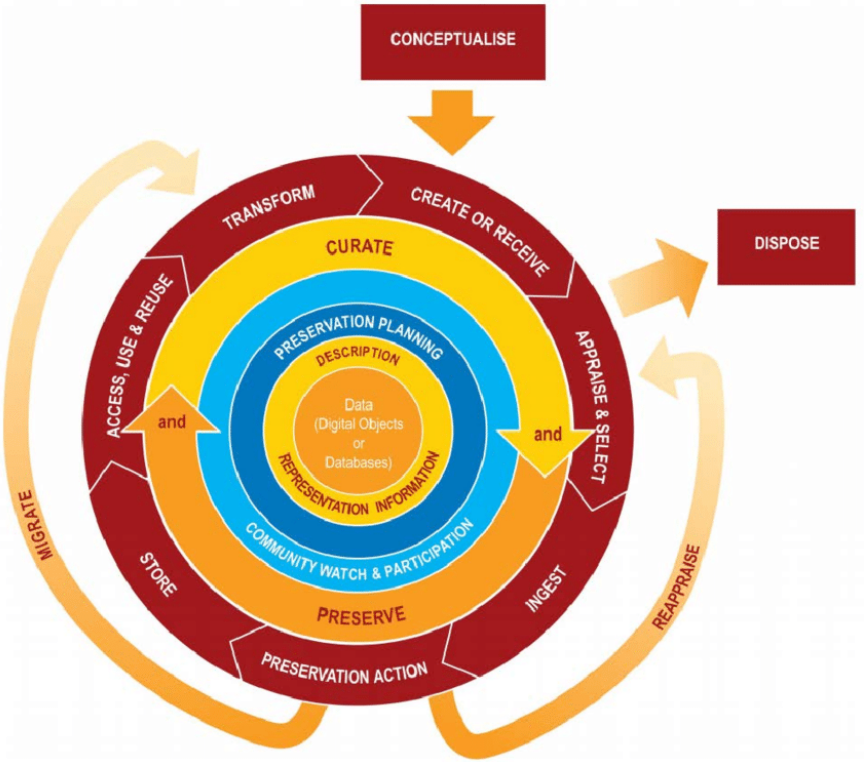
- There are several roles in the data publishing process: *Author*, *Curator* and *Consumer*.
- Digital curation and data preservation are ongoing processes, requiring considerable thought and the investment of adequate time and resources. You must be aware of, and undertake, actions to promote curation and preservation throughout the data lifecycle
- Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle
- The active management of research data reduces threats to their long-term research value and mitigates the risk of digital obsolescence. Meanwhile, curated data in trusted digital repositories may be shared among the wider research community
- As well as reducing duplication of effort in research data creation, curation enhances the long-term value of existing data by making it available for further high quality research

Data Life Cycle – Digital Curation Centre



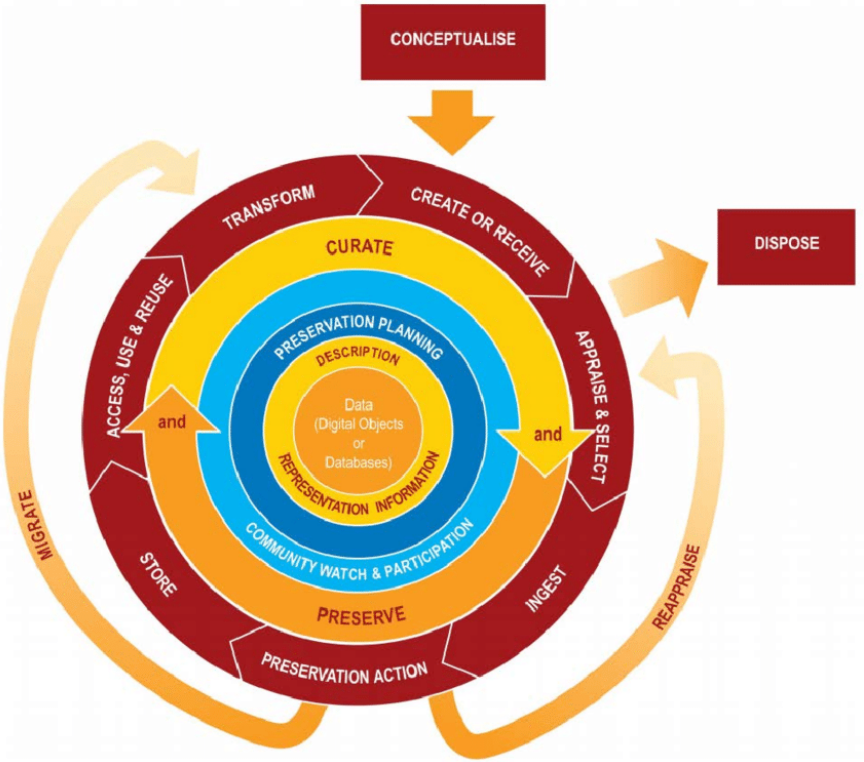
- Lifecycle management of digital materials is necessary to ensure their continuity
- The DCC Curation Lifecycle Model has been developed as a generic, curation-specific, tool which can be used, in conjunction with relevant standards, to plan curation and preservation activities to different levels of granularity
- The DCC will use the model:
  - as a training tool for data creators, data curators and data users;
  - to organise and plan their resources;
  - to help organisations identify risks to their digital assets and plan management strategies for their successful curation

DLC DCC – Data



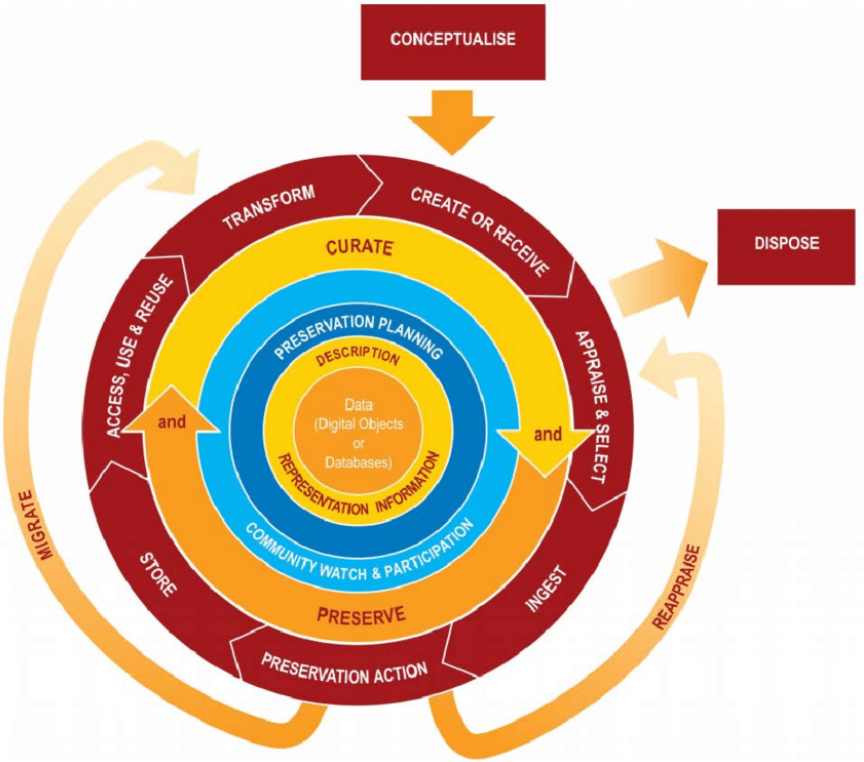
- Data, any information in binary digital form, is at the centre of the Curation Lifecycle
- Digital Objects
  - Simple Digital Objects are discrete digital items; such as textual files, images or sound files, along with their related identifiers and metadata.
  - Complex Digital Objects are discrete digital objects, made by combining a number of other digital objects, such as websites
- Databases
  - Structured collections of records or data stored in a computer system

DLC DCC – Full Lifecycle Actions



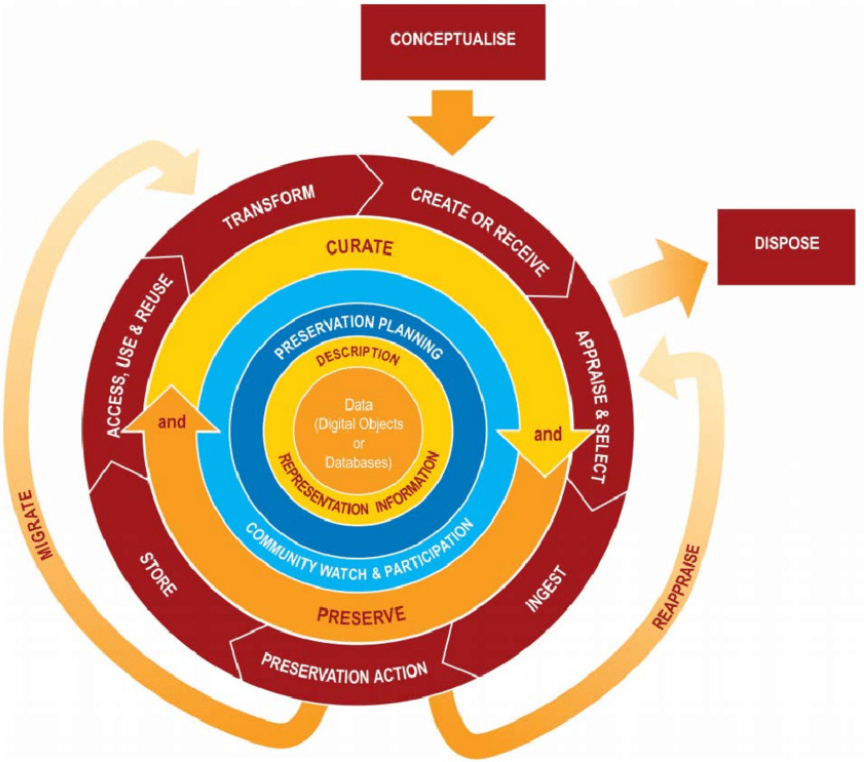
- Description and Representation Information
  - Assign administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long term. Collect and assign representation information required to understand and render both the digital material and the associated metadata.
- Preservation Planning
  - Plan for preservation throughout the curation lifecycle of digital material. This would include plans for management and administration of all curation lifecycle actions.
- Community Watch and Participation
  - Maintain a watch on appropriate community activities, and participate in the development of shared standards, tools and suitable software.

DLC DCC – Sequential Actions (1)



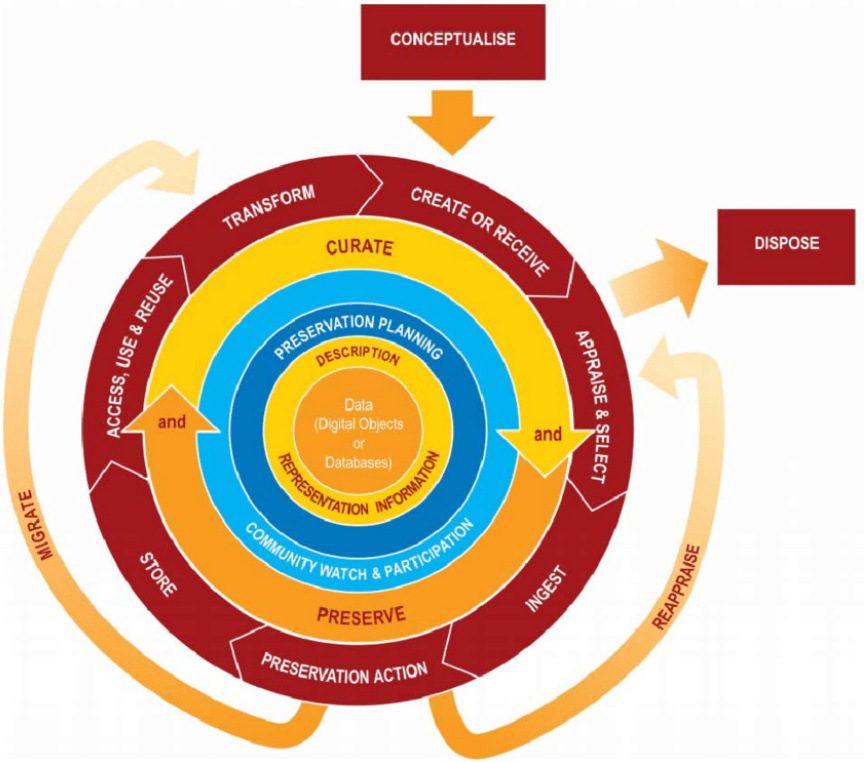
- Conceptualise
  - Conceive and plan the creation of data, including capture method and storage options.
- Create and Receive
  - Create data including administrative, descriptive, structural and technical metadata. Preservation metadata may also be added at the time of creation.
  - Receive data, in accordance with documented collecting policies, from data creators, other archives, repositories or data centres, and if required assign appropriate metadata.
- Appraise and Select
  - Evaluate data and select for long-term curation and preservation. Adhere to documented guidance, policies or legal requirements.
- Ingest
  - Transfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements.

DLC DCC – Sequential Actions (2)



- Preservation Action
  - 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
  - Store the data in a secure manner adhering to relevant standards.
- Access, Use and Reuse
  - 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.
- Transform
  - 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.

DLC DCC – Occasional Actions

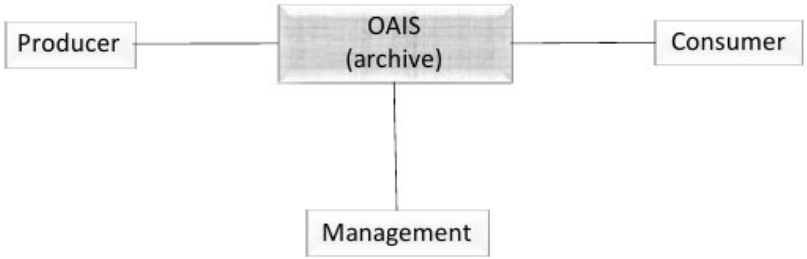


- Dispose
  - Dispose of data, which has not been selected for long-term curation and preservation in accordance with documented policies, guidance or legal requirements. Typically data may be transferred to another archive, repository, data centre or other custodian. In some instances data are destroyed. The data's nature may, for legal reasons, necessitate secure destruction.
- Reappraise
  - Return data which fails validation procedures for further appraisal and reselection.
- Migrate
  - Migrate data to a different format. This may be done to accord with the storage environment or to ensure the data's immunity from hardware or software obsolescence.

Open Archival Information System

- An Open Archival Information System (OAIS) is an Archive, consisting of an organization, which may be part of a larger organization, of people and systems that has accepted the responsibility to preserve information and make it available for a Designated Community.
- The information being maintained has been deemed to need Long Term Preservation, even if the OAIS itself is not permanent. Long Term is long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community. Long Term may extend indefinitely.
- The OAIS model be applicable to any Archive.
- A conformant OAIS Archive may provide to users additional services that are beyond those required of an OAIS.
- The OAIS Reference Model is the standard ISO 14721:2012.

OAIS – Environment



- Outside the OAIS are Producers, Consumers, and Management.
- **Producer** is the role played by those persons, or client systems, which provide the information to be preserved.
- **Management** is the role played by those who set overall OAIS policy as one component in a broader policy domain, for example as part of a larger organization.
- **Consumer** is the role played by those persons, or client systems, that interact with OAIS services to find and acquire preserved information of interest. A special class of Consumers is the Designated Community. The Designated Community is the set of Consumers who should be able to understand the preserved information.

OAIS – Information

- A person, or system, can be said to have a **Knowledge Base**, which allows that person or system to understand received information.
- **Information** is any type of knowledge that can be exchanged. In an exchange, it is represented by **Data Object**.
- **Representation Information** is the information that maps a Data Object into more meaningful concepts.
- **Information Object** is a Data Object together with its Representation Information.

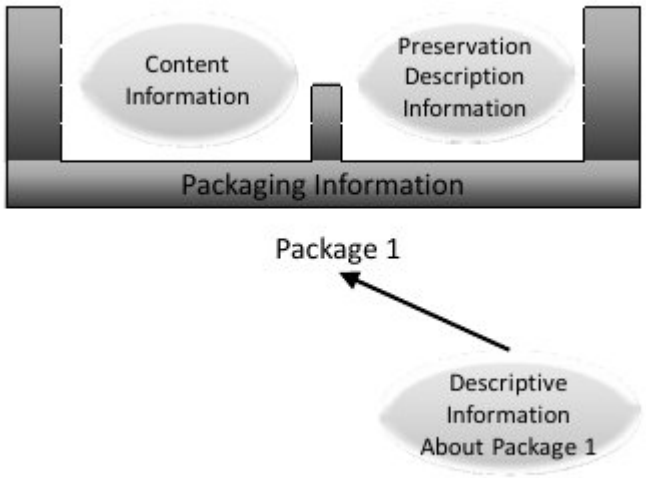
- An OAIS must identify and understand clearly the Data Object and its Representation Information
- An OAIS must collect all the relevant Representation Information or reference its existence in another trusted or partner

OAIS – Information Package

- Every submission of information to an OAIS by a Producer, and every dissemination of information to a Consumer, occurs as one or more discrete transmissions called **Information Package**.
- An Information Package is a conceptual container of two types of information called **Content Information** and **Preservation Description Information (PDI)**. The resulting package is viewed as being discoverable by virtue of the **Descriptive Information**.
  - The Content Information is a set of information that is the original target of preservation or that includes part or all of that information. It is an Information Object composed of its Content Data Object and its Representation Information.
  - The PDI is divided into five types of preserving information called Provenance, Context, Reference (e.g. an unique identifier), Fixity (e.g. a checksum) and Access Rights
  - The Descriptive Information provides metadata to support the finding, ordering, and retrieving of OAIS information holdings by Consumers.

OAIS – Information Package

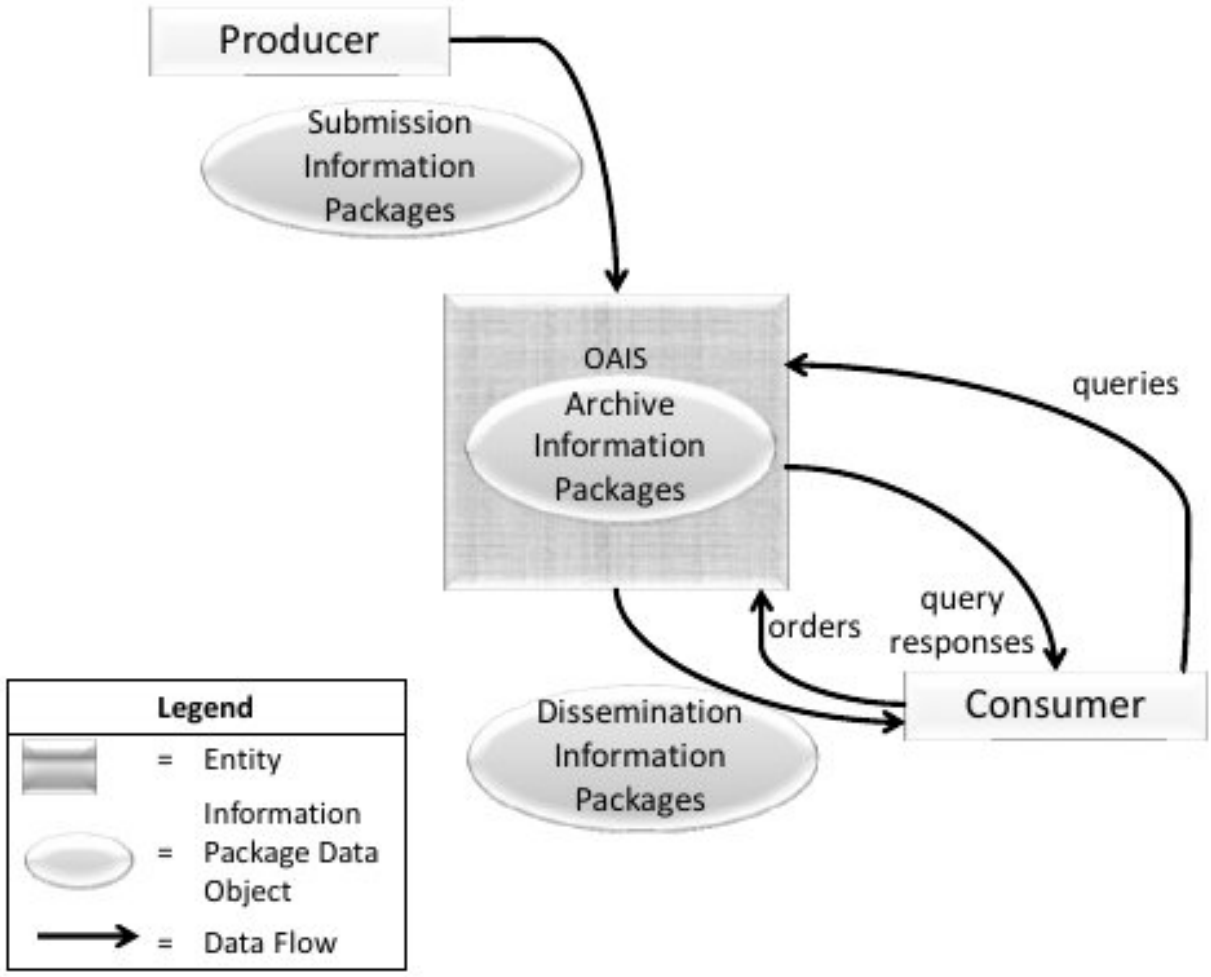
- Every submission of information to an OAIS by a Producer, and every dissemination of information to a Consumer, occurs as one or more discrete transmissions called **Information Package**.
- An Information Package is a conceptual container of two types of information called **Content Information** and **Preservation Description Information (PDI)**. The resulting package is viewed as being discoverable by virtue of the **Descriptive Information**.
  - The Content I… s the original target of preservation o… ation. It is an Information Object compos… Representation Information.
  - The PDI is div… mation called Provenance, C… ntifier), Fixity (e.g. a checksum) an…
  - The Descriptiv… pport the finding, ordering, and … s by Consumers.



OAIS – Information Package Variants

- In principle Information Packages submitted to an OAIS by a Producer, preserved by an OAIS, and disseminated to a Consumer may be different.
  - Insufficient Representation Information or PDI when submitted
  - Information Packages are organized differently when stored or disseminated providing additional Content Information
- The **Submission Information Package (SIP)** is that package that is sent to an OAIS by a Producer. Its form and detailed content are typically negotiated between the Producer and the OAIS. Most SIPs will have some Content Information and some PDI.
- Within the OAIS one or more SIPs are transformed into one or more **Archival Information Packages (AIPs)** for preservation. The AIP has a complete set of PDI for the associated Content Information. The AIP may also contain a collection of other AIPs.
- In response to a request, the OAIS provides all or a part of an AIP to a Consumer in the form of a **Dissemination Information Package (DIP)**. The DIP may also include collections of AIPs, and it may or may not have complete PDI. The Packaging Information will necessarily be present in some form so that the Consumer can clearly distinguish the information that was requested. Depending on the dissemination media and Consumer requirements, the Packaging Information may take various forms.

OAIS – Interactions



OAIS – Management Interaction

- Management provides the OAIS with its charter and scope.
- Some examples of typical interactions between the OAIS and Management include:
  - Management is often the primary source of funding for an OAIS and may provide guidelines for resource utilization (personnel, equipment, facilities).
  - Management will generally conduct some regular review process to evaluate the OAIS performance and progress toward Long Term goals, and assess the risks to which the OAIS and its holdings are exposed.
  - Management determines, or at least endorses, pricing policies, as applicable, for OAIS services.
  - Management participates in conflict resolution involving Producers, Consumers and OAIS internal administration.

OAIS – Producer Interaction

- The first contact between the OAIS and the Producer is a request that the OAIS preserve the data products created by the Producer establishing a **Submission Agreement** involving the Producer and the Management.
- Within the Submission Agreement, one or more **Data Submission Sessions** are specified.

OAIS – Consumer Interaction

- The Consumer establishes an **Order Agreement** with the OAIS for information.
- The Order Agreement may span any length of time, and under it one or more **Data Dissemination Sessions** may take place.
- The Consumer will establish a **Search Session** with the OAIS. During this Search Session the Consumer will use the OAIS **Finding Aids** that operate on **Descriptive Information**, or in some cases on the AIPs themselves, to identify and investigate potential holdings of interest.

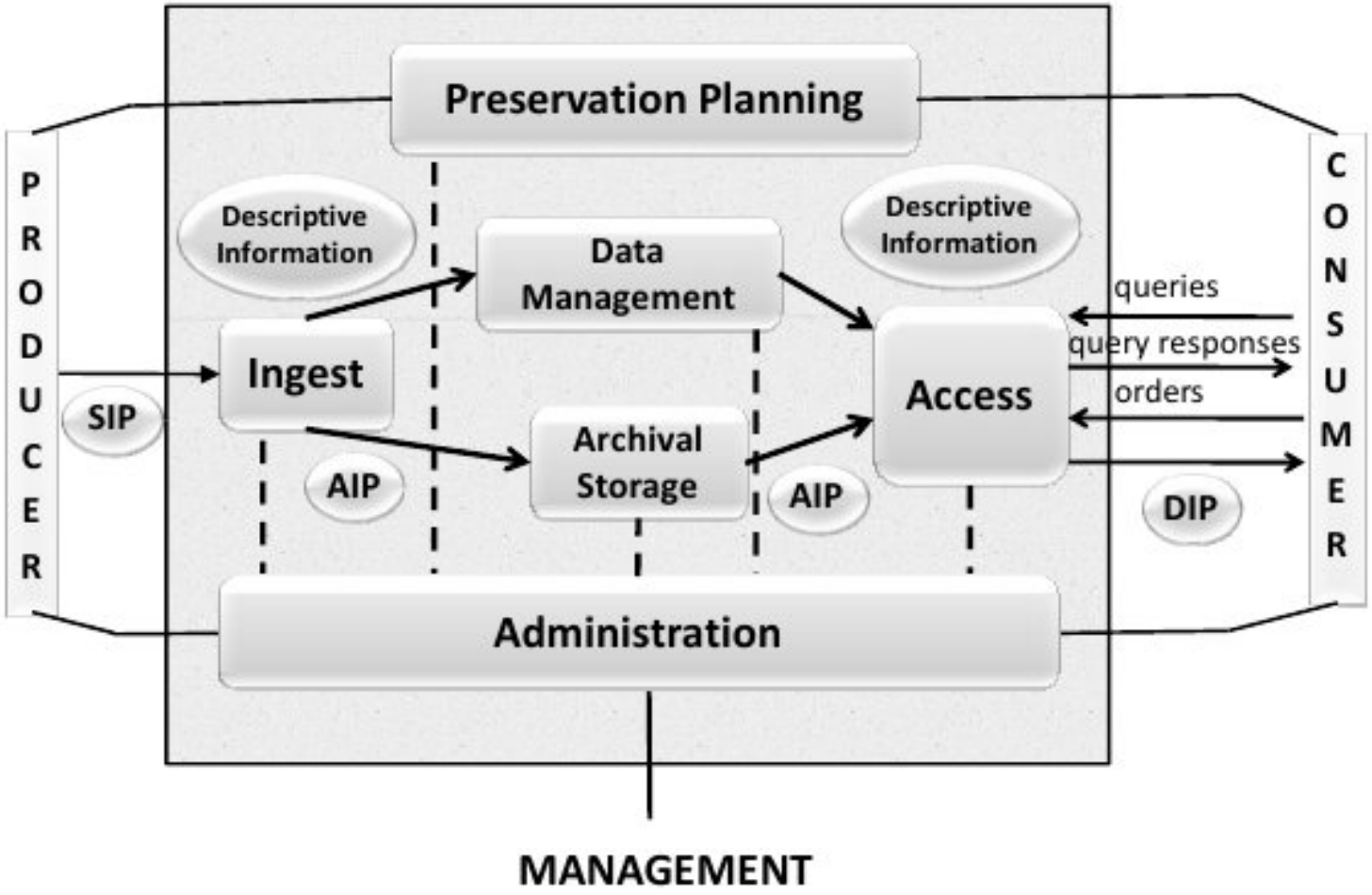
OAIS – Responsibilities (1)

- Negotiate for and accept appropriate information from information Producers.
- Obtain sufficient control of the information provided to the level needed to ensure Long Term Preservation. This include:
  - Assume sufficient control over the Content Information and Preservation Description Information so that it is able to preserve it for the Long Term
  - Copyright implications, intellectual property and other legal restrictions
  - Authority to modify Content Information
  - Agreements with external organizations
- Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided, thereby defining its Knowledge Base.

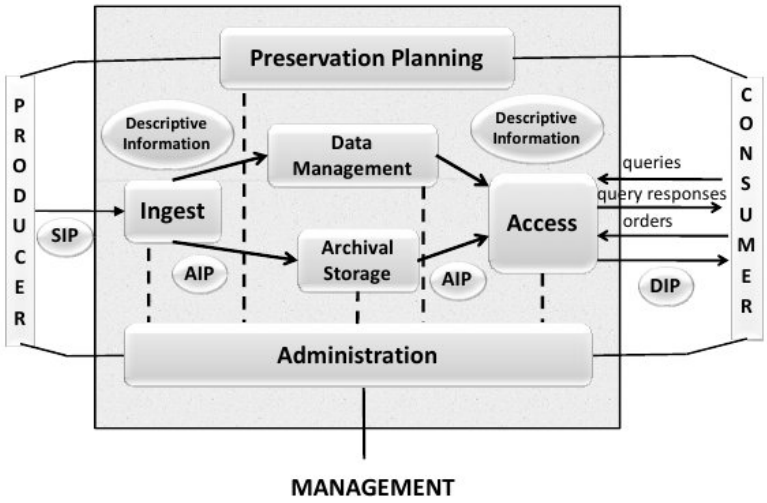
OAIS – Responsibilities (2)

- Ensure that the information to be preserved is Independently Understandable to the Designated Community. In particular, the Designated Community should be able to understand the information without needing special resources such as the assistance of the experts who produced the information.
- Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, including the demise of the Archive, ensuring that it is never deleted unless allowed as part of an approved strategy. There should be no ad-hoc deletions.
- Make the preserved information available to the Designated Community and enable the information to be disseminated as copies of, or as traceable to, the original submitted Data Objects with evidence supporting its Authenticity.

OAIS – Functional Model (1)

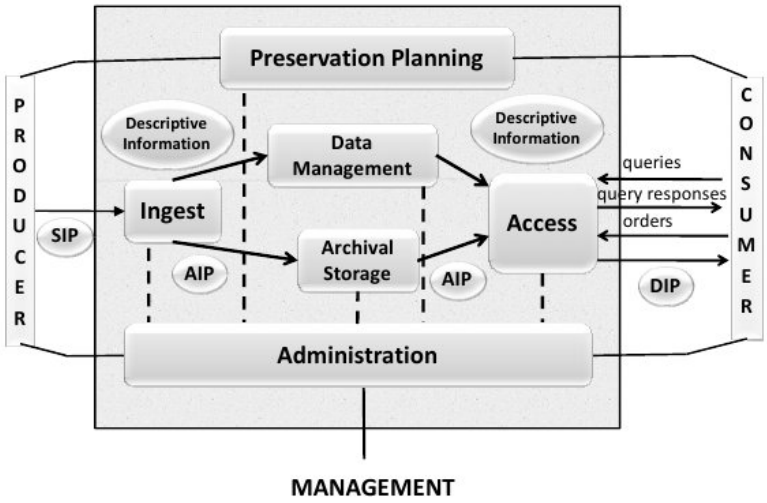


OAIS – Functional Model (2)



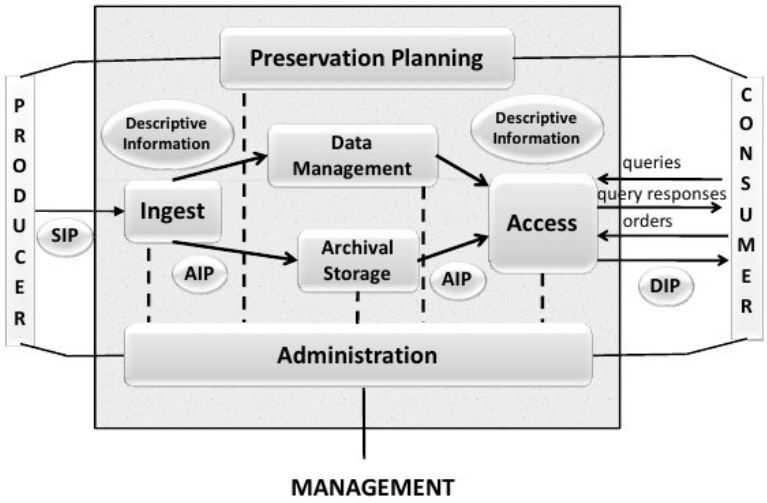
- The **Ingest** provides the services and functions to accept SIPs from Producers and prepare the contents for storage and management. It includes:
  - receiving SIPs
  - performing quality assurance on SIPs
  - generating AIPs
  - extracting Descriptive Information from the AIPs for inclusion in the Archive database
  - coordinating updates to Archival Storage and Data Management

OAIS – Functional Model (3)



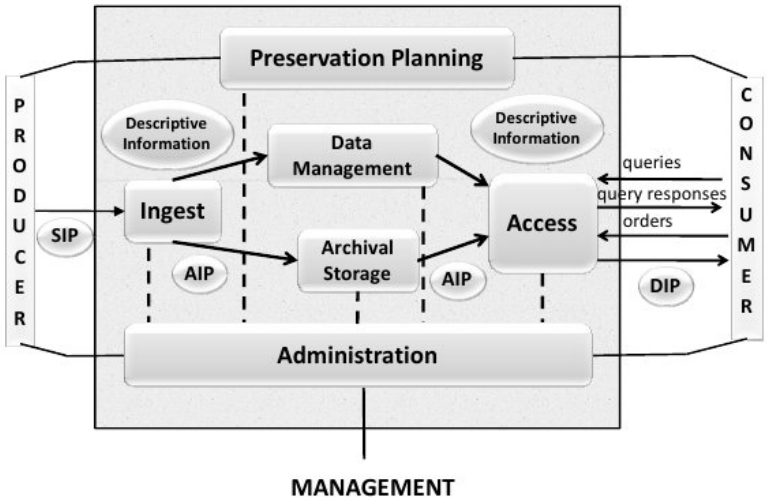
- The **Archival Storage** provides the services and functions for the storage, maintenance and retrieval of AIPs. It includes:
  - receiving AIPs from Ingest and adding them to permanent storage
  - managing the storage hierarchy
  - performing routine and special error checking
  - providing disaster recovery capabilities
  - providing AIPs to Access to fulfill orders

OAIS – Functional Model (4)



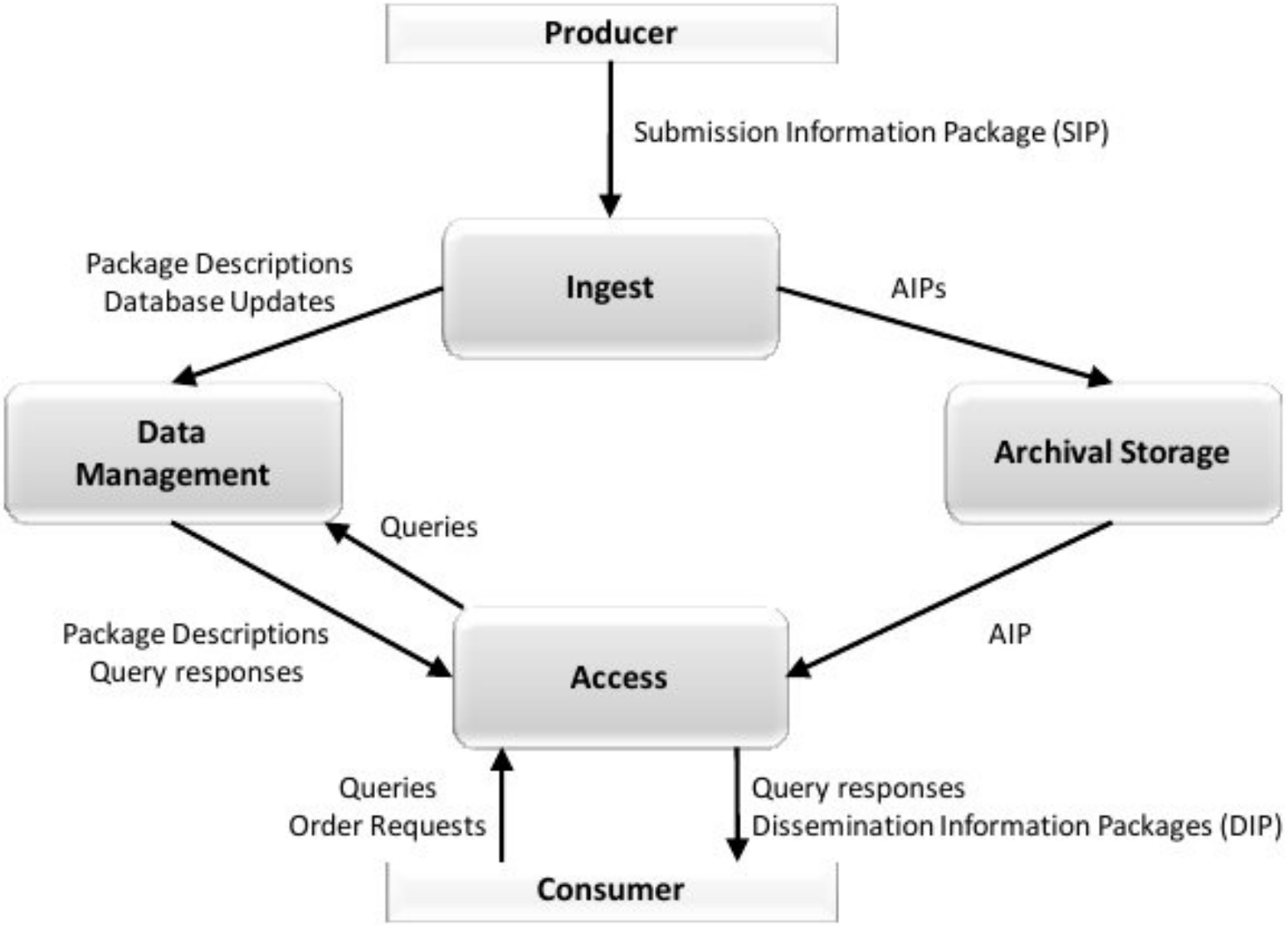
- The **Data Management** provides the services and functions for populating, maintaining, and accessing both Descriptive Information and administrative data used to manage the Archive. It includes:
  - administering the Archive database functions
  - performing database updates
  - performing queries on the data management data to generate query responses

OAIS – Functional Model (5)



- The **Access** provides the services and functions that support Consumers in determining the existence, description, location and availability of information stored in the OAIS, and allowing Consumers to request and receive information products. It includes:
  - communicating with Consumers to receive requests
  - applying controls to limit access to specially protected information
  - coordinating the execution of requests to successful completion
  - generating responses (DIPs, query responses, reports) and delivering the responses to Consumers

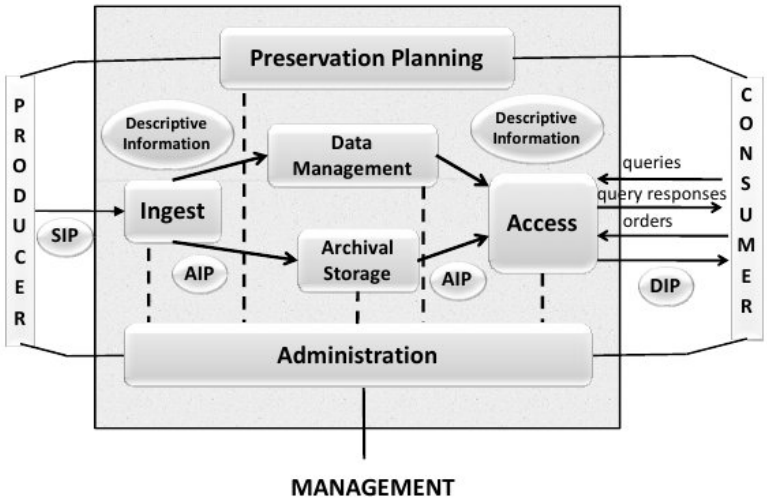
OAIS – Information Package Transformations



OAIS – Information Package Transformations

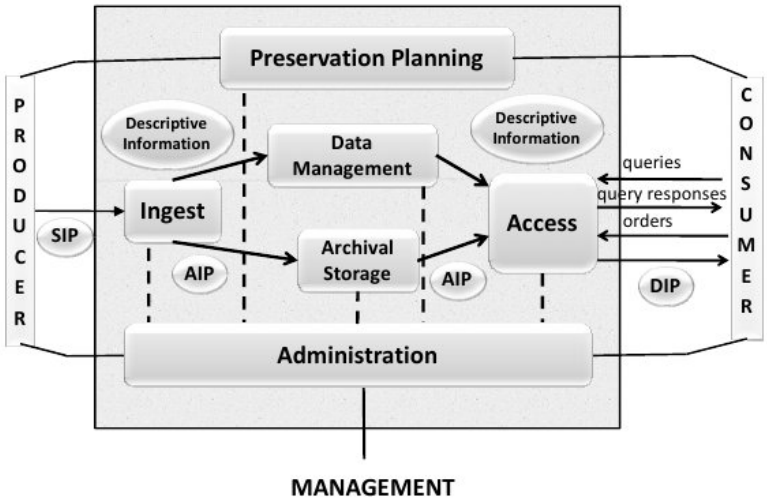
- The mapping between SIPs and AIPs is not one-to-one. Here are some examples:
  - One SIP—One AIP: A government agency is ready to Archive its electronic records from the previous fiscal year. All of the year's records are placed onto magnetic tapes that are submitted as one SIP. The Archive stores the tapes together as a single AIP.
  - Many SIPs—One AIP: A satellite sensor makes observations of the Earth over a period of one year. Every week all of the latest sensor data are submitted to the Archive as a SIP. The Archive has a single AIP containing all of the sensor's observations for the year. Ingest merges the Content Information from each weekly SIP into a specified file/files in Ingest persistent storage. The PDI data for the AIP is sent after the last sensor data for the year has been received. After all of the weekly SIPs and the SIP containing the PDI have arrived, Ingest processes the AIP.
  - One SIP—Many AIPs: A company submits financial records to an Archive as one SIP. The Archive chooses to store this information as two AIPs: one that contains public information and the other that contains sensitive information. This makes it easier for the Archive to manage access to the information.
  - Many SIPs—Many AIPs: An oil and gas company collects information on its wells. Every year it submits SIPs containing all of the well status information for one well to an Archive. The Archive maintains one AIP for each oil or gas field and breaks out the information on each well to the proper AIP based upon its geographic coordinates.

OAIS – Functional Model (6)



- The **Preservation Planning** provides the services and functions for monitoring the environment of the OAIS, providing recommendations and preservation plans to ensure that the information stored in the OAIS remains accessible to, and understandable by, the Designated Community over the Long Term. It includes:
  - evaluating the contents of the Archive and periodically recommending archival information updates
  - recommending the migration of current Archive holdings
  - developing recommendations for Archive standards and policies, providing periodic risk analysis reports, and monitoring changes in the technology environment and in the Designated Community's service requirements and Knowledge Base

OAIS – Functional Model (7)



- The **Administration** provides the services and functions for the overall operation of the Archive system. It includes:
  - soliciting and negotiating submission agreements with Producers
  - auditing submissions to ensure that they meet Archive standards
  - maintaining configuration management of system hardware and software
  - monitoring and improving Archive operations to inventory, report on, and migrate/update the contents of the Archive
  - establishing and maintaining Archive standards and policies, providing customer support, and activating stored requests

OAIS – Preservation Perspectives

- Preservation practices must face the fast-changing nature of the computer industry and the ephemeral nature of electronic data storage media, operating systems, etc.
- **Digital Migration** of an AIP can include:
  - copying Content Data Object or Representation Information bits to new media
  - altering or adding to Content Data Object or Representation Information bits
  - altering or adding to PDI bits
  - altering or adding to operational software whose role is essential to Content Information preservation (i.e., it is part of Representation Information)
  - altering or adding to the bits that make up the AIP's Packaging Information
- Digital Migration which does not change the bit sequences
  - Refreshment: AIPs are moved to new media instance replacing the old media instance
  - Replication: like Refreshment but may require changes to the Archival Storage mapping infrastructure
- Digital Migration which changes the bit sequences
  - Repackaging: some changes in the bits pf Packaging Information are required
  - Transformation: some changes in the Content Information or PDI are required

OAIS – Distinguishing Versions

- An AIP may, in some environments, be subject to upgrading or improvement over time.
- This is not a Digital Migration in that the intent is not to preserve information, but to increase or improve it.
- This type of AIP change may be referred to as creating a new AIP Edition.
  - The AIP Edition may or may not be viewed as a replacement for the source AIP, but it may be of historical interest to retain the previous AIP.
  - This also results in a new AIP ID with the same impacts on Associated Descriptions and Access Aids as a Digital Migration Transformation.
- An OAIS may also find it convenient to provide an AIP that is derived from an existing AIP.
  - It may do this by extracting some information, or by aggregating information from multiple AIPs, to better serve Consumers.
  - This type of resulting AIP may be referred to as a Derived AIP.
  - It does not replace any of the AIPs that it was derived from and it is not a result of a Digital Migration.
  - This also results in a new AIP ID and a new Associated Descriptions.
  - This may also require updates to, or new, Access Aids depending on how they have been implemented.

OAIS – Archive Interoperability

- Consumer, Producers and Manager may wish to have
  - common finding aids to aid in locating information across several Archives
  - a common Package Description schema for access or a common DIP schema
  - a single global access site
  - a common SIP schema for submission to different Archives
  - a single depository for all their products
  - cost reduction through sharing of expensive hardware, software, and preservation efforts
  - increasing the uniformity and quality of interactions with several Archives
- Independent Archives are motivated by local concerns with no management or technical interaction among them.
- Federated Archives are based on standards agreements among two or more Archives, and must provide
  - Unique AIP Names for each AIP in the Federation.
  - User Authentication and Access Management for global users.

References

- Digital Curation Centre (DCC) Lifecycle Model
  http://www.dcc.ac.uk/resources/curation-lifecycle-model
- The Consultative Committee for Space Data Systems (CCSDS)
  https://public.ccsds.org
- Open Archival Information System (OAIS) Reference Model
  https://public.ccsds.org/Pubs/650x0m2.pdf
- Telescopio Nazionale Galileo (TNG) archive
  http://archives.ia2.inaf.it/tng
- Mikulski Archive for Space Telescopes (MAST) archive
  https://archive.stsci.edu