

An Organisational Model for Digital Archive Centres

Claude HUC (CNES – French Space Agency)

ur context: Long-term archival of space data

igital preservation: a problem shared by everyone

he Organizational Model proposed

- Overview
- Description of each service: functions, responsibilities, external interfaces, s required

onclusions and References

Space Data

early always produced in digital format

diversity and wealth of scientific information - numerous subjects:
astronomy, earth observation, space physics, fundamental physics
ology, etc.

- Observations from scientific instruments
- Processing results of these observations, calibrations
- Results from modelling processing
- Documents describing instruments, experimental processes
- Metadata
- Publications

- Diversity of representation formats
 - Text documents, sets of digital values, images, vectorized graphics, video, sound, multimedia documents, etc.
- Very large volumes
- Complexity
- Geographical dispersion

Zoom in on the problem of preservation facing all sectors of activity





Why preserve space data?

Because the observations are often unique and cannot be reproduced

- A comet or a solar flare, for example

Because scientific analysis often involves observations over very long periods of time

- For example, studying changes in the earth's climate

Because these observations could, in the future, be processed using new analysis methods not currently available

etc.

Lessons learnt at CNES

Space data in digital format for the last 40 years
Accelerated obsolescence of technologies since 1990
Backup program conducted from 1995 to 2000

- Justified by the expected discontinuation of magnetic tape storage technology
- Most data has been saved, but useful scientific observations have nevertheless been lost
- ==> Some damaged media, but even more often the description of the information is incomplete, inaccurate or even unavailable

Large text documents entered on word processor in 1985

- *Entered under MS Word again in 1990 (Word 2) then entered again under MS Word in 1997 (Word 95)*

Compatibility chain broken in less than 10 years

Lessons learnt at CNES

Progressive elaboration of pragmatic solutions

Large number of functions

Large diversity of needed skills :

- Physical preservation of files, storage media
- Representation information formats
- Knowledge and understanding of the archived information
- Technologies for search and retrieval

relative paralysis in front of the growing obsolescence of technologies

- Impacts on the costs
- How to deal with such a situation ?

Lessons learnt at CNES

an excessive focus on technical problems (media, data formats..)

the separation between

- The cases for which we have valid organisational and technical approaches propose
- The cases for which we don't have reliable solution :
 - ❖ Example : the emulation ways are always part of the research activity and not p acceptable operational possibilities for organization in charge of information preservation



The problem: reduction of time scales

Problems encountered:

- ❖ *office documents*: ==> **compatibility chain broken in less than 10 years**
- ❖ More recent scientific documents (1995!!) but for which all mathematical formulas had to be re-typed.
- ❖ Scientific observations stored on magnetic tapes saved at the last minute

Technological change accelerating since the 1990's

- This trend shows no signs of slackening off, quite the contrary (at least 5 versions of MS Word for Windows since 1995).

the preservation of a digital document saved 10 years ago, or even less, may already be vulnerable.

Who is concerned today?



Almost everyone:

- Administration: public records, etc.
- Health care sector
- Pension funds
- Industry: oil, aeronautics, etc.
- Scientific research, the space sector
- Defence
- Nuclear sector
-as well as private individuals.

The French PIN Group

N: *Pérennisation des Informations Numériques* (Preservation of digital information)

Inter-organizational work group set up in 2000

- within the *Aristote* non-profit association (<http://pin.cnes.fr>)

Fields of activity:

- Terminology and standards
- Information representation formats
- Archiving systems
- Organizational: roles of the participants
- Information to be archived: classification, formats, life cycle, etc.
- Legal issues

Actions

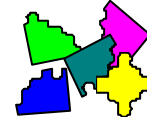
- Methodology used to evaluate formats for preservation
- Creation of global, comprehensive and practical training



Participants in the PIN Group

French National Archives, Centre of Contemporary Archives (*Direction des Archives de France, Centre des Archives Contemporaines*)
French National Library (*BnF: Bibliothèque nationale de France*)
French Atomic Energy Commission (*CEA*)
National Computer Centre of Higher Education (*CINES: Centre Informatique National de l'Enseignement Supérieur*)
French Agricultural Research Centre for International Development (*CIRAD: Centre de coopération Internationale en Recherche Agronomique pour le Développement*)
France Télécom Research and Development
Groupe Médéric (pension funds)
French Institute of Research on the Utilization of the Sea (*IFREMER: Institut Français de Recherche pour l'Exploitation de la MER*)
French National Audiovisual Institute (*INA: Institut National de l'Audiovisuel*)
National Geographical Institute (*IGN: Institut Géographique National*)
INRIA
Ministries of Agriculture, Public Works, Justice and Defence
City of Paris
Institut Pasteur

Vulnerability: numerous causes



Technical factors

- Obsolescence of storage technologies, software and systems
- Dependency between data created and the creation environment
- Data or documents not described

Organizational and financial factors

- Preservation of information is an activity in its own right.
- **Need to review work organisation, sharing of responsibility**
- **setting up the right skills at the right place**



Normative, legal, industrial, psychological factors, related to lack of basic training...



Understand the problem posed in order to solve it

This is the purpose of the 'Reference Model for an Open Archival Information System (OAIS)' Issue 1. January 2002

<http://www.ccsds.org/CCSDS/recommandreports.jsp#interchange>
CCSDS reference: CCSDS 650.0-B-1 (free)

<http://www.iso.org/> ISO reference: ISO 14721:2002 (206 Swiss Francs...)

Translation in progress, conducted jointly between CNES and the BnF

detailed analysis, definition of concepts, a functional model and an information model to **understand** all aspects involved in archiving information in digital format

OAIS: What is an archive?

structure whose aim is to preserve information for access and use by a **Defined Community of Users**.

- Data preservation
- Long-term access to data
- Along with data, preservation of all information required for its comprehension and use

Definition taken from ISO standard 14721:2002

Archival is not:

- a copy or a system backup
- data filed away permanently when it is not expected to be used any more

OAIS: Data and information

Information

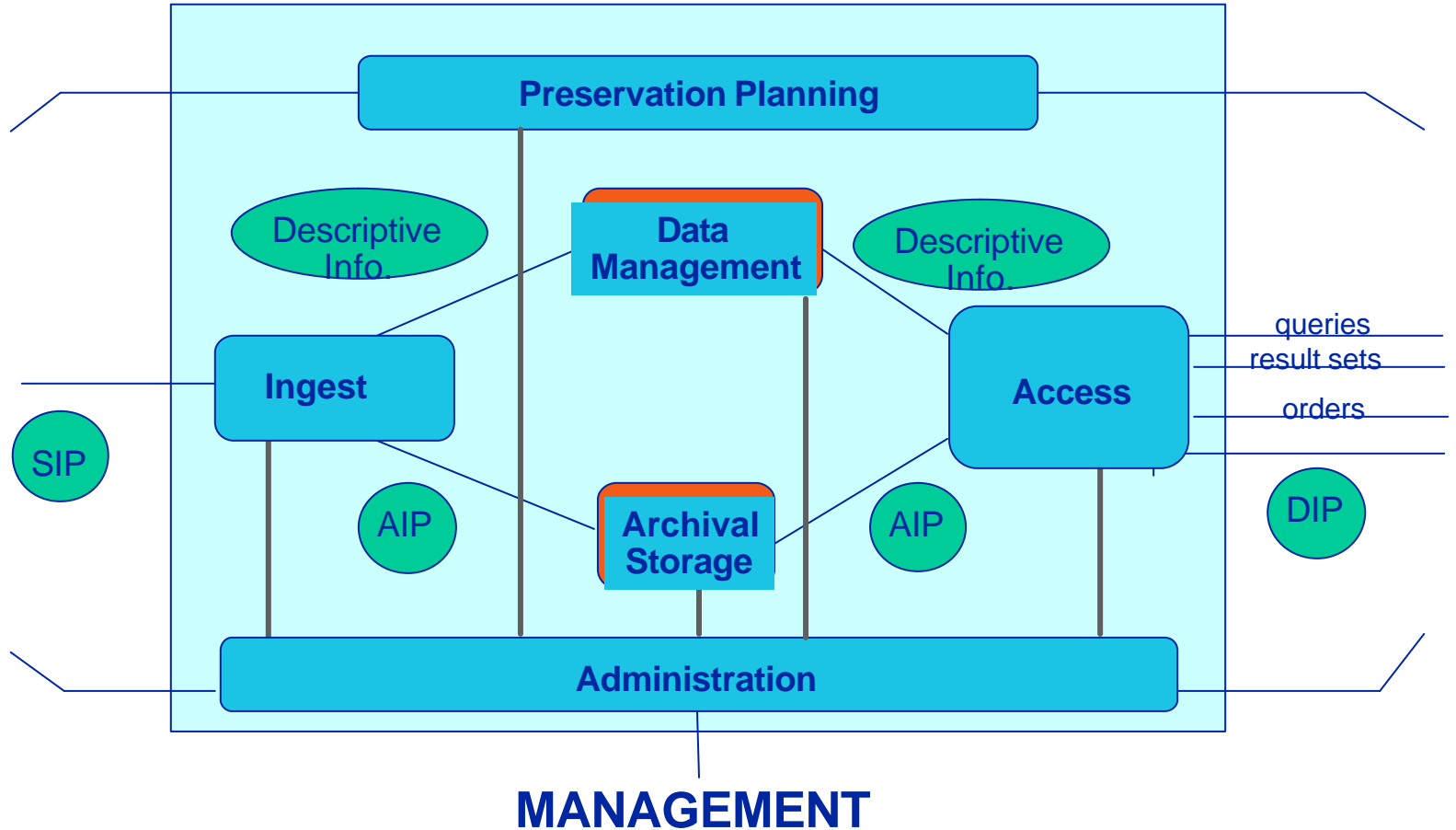
- any type of knowledge which can be exchanged
- independent of the format (i.e. physical or digital) used to represent this information

Data: information representation formats

OAIS: Functional Entities

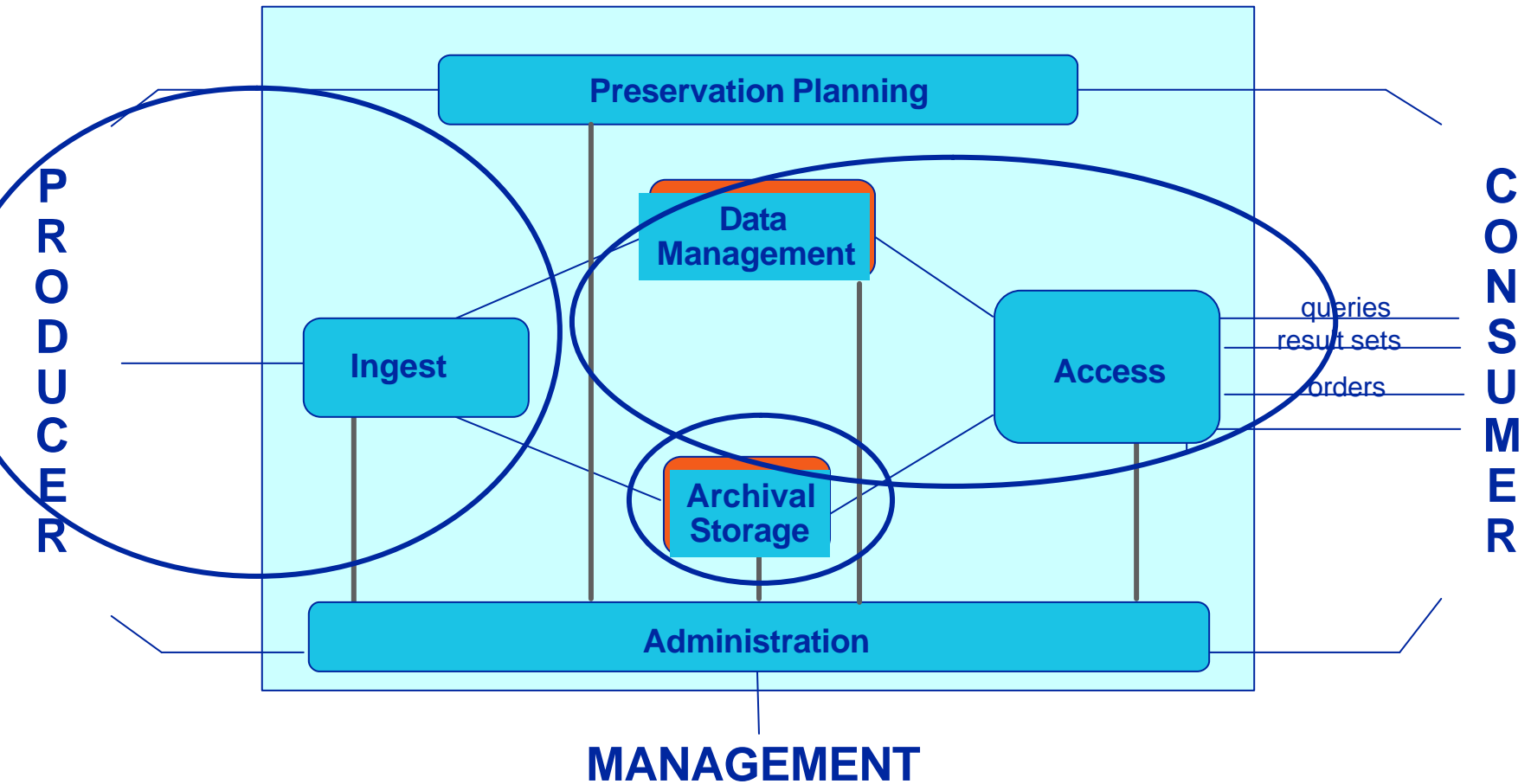
PRODUCER

CONSUMER



- SIP = Submission Information Package
- AIP = Archival Information Package
- DIP = Dissemination Information Package

Search for a practical structure





OAIS: Functional Entities

With a very large number of functions and sub-functions we have to deal with

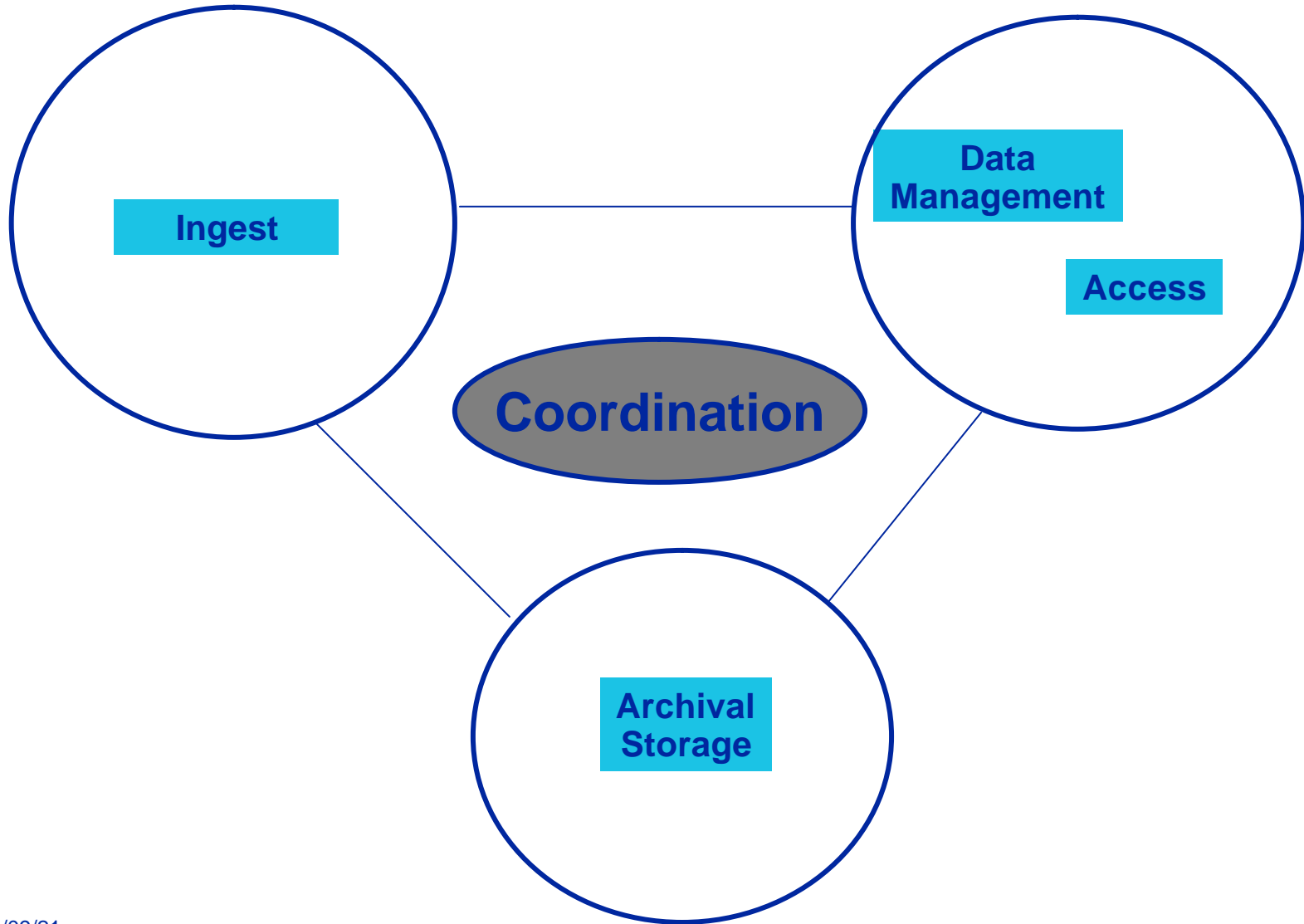
We do not know clearly how to start to solve the problem

The existing organization for traditional (non digital) archives is not really appropriate for digital information

In this framework, we propose an organization split in autonomous services – each service is in charge of precise functions with well defined interfaces with other services : reduce the size of the problem should help us to solve it

- Link the abstract and global OAIS approach to the pragmatic practices

Three coordinated services



CONSUMER

Service?

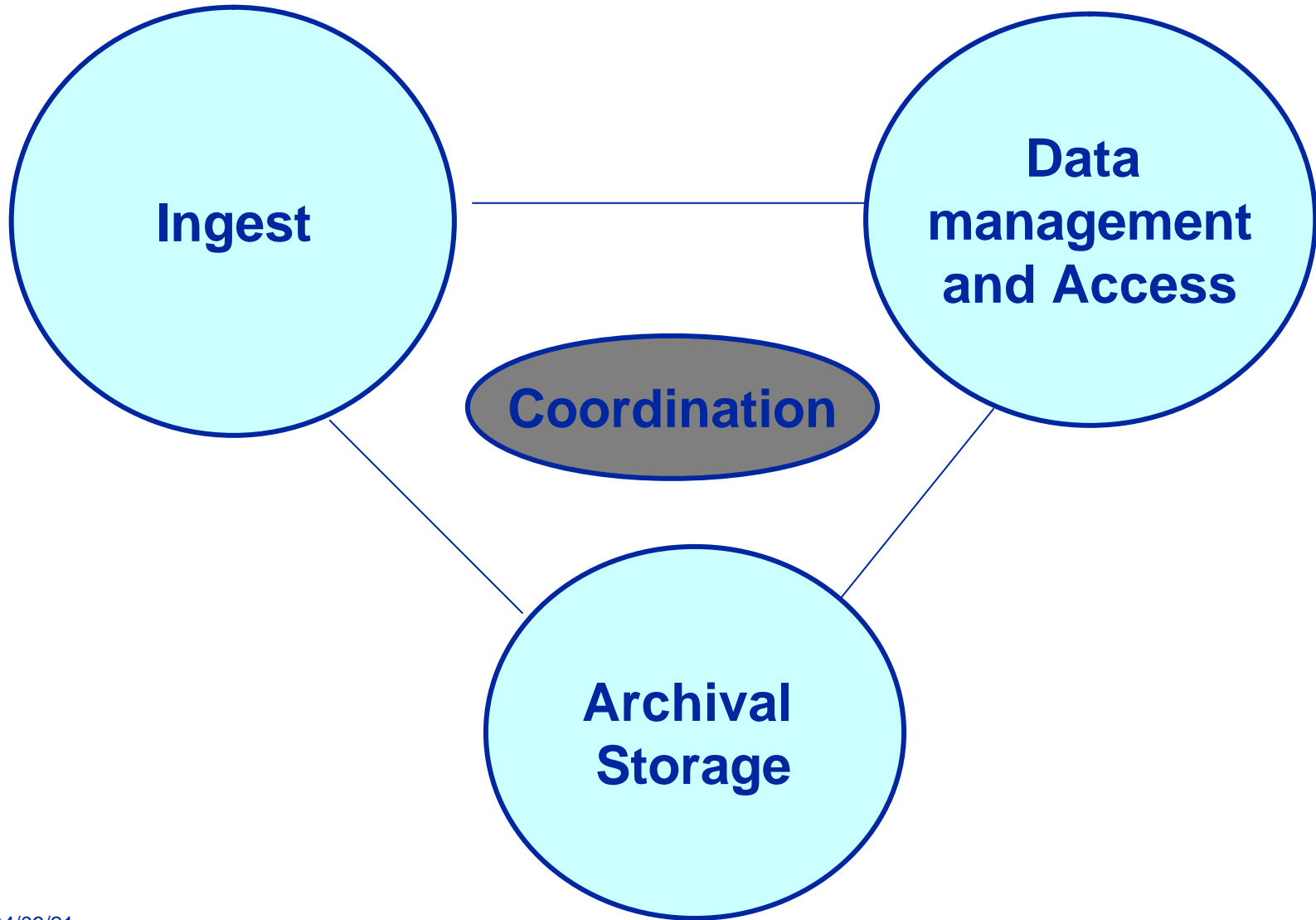
Service' is taken to mean an administrative unit consisting of people, technical facilities and resources given a clear assignment

For each service proposed, it is necessary to:

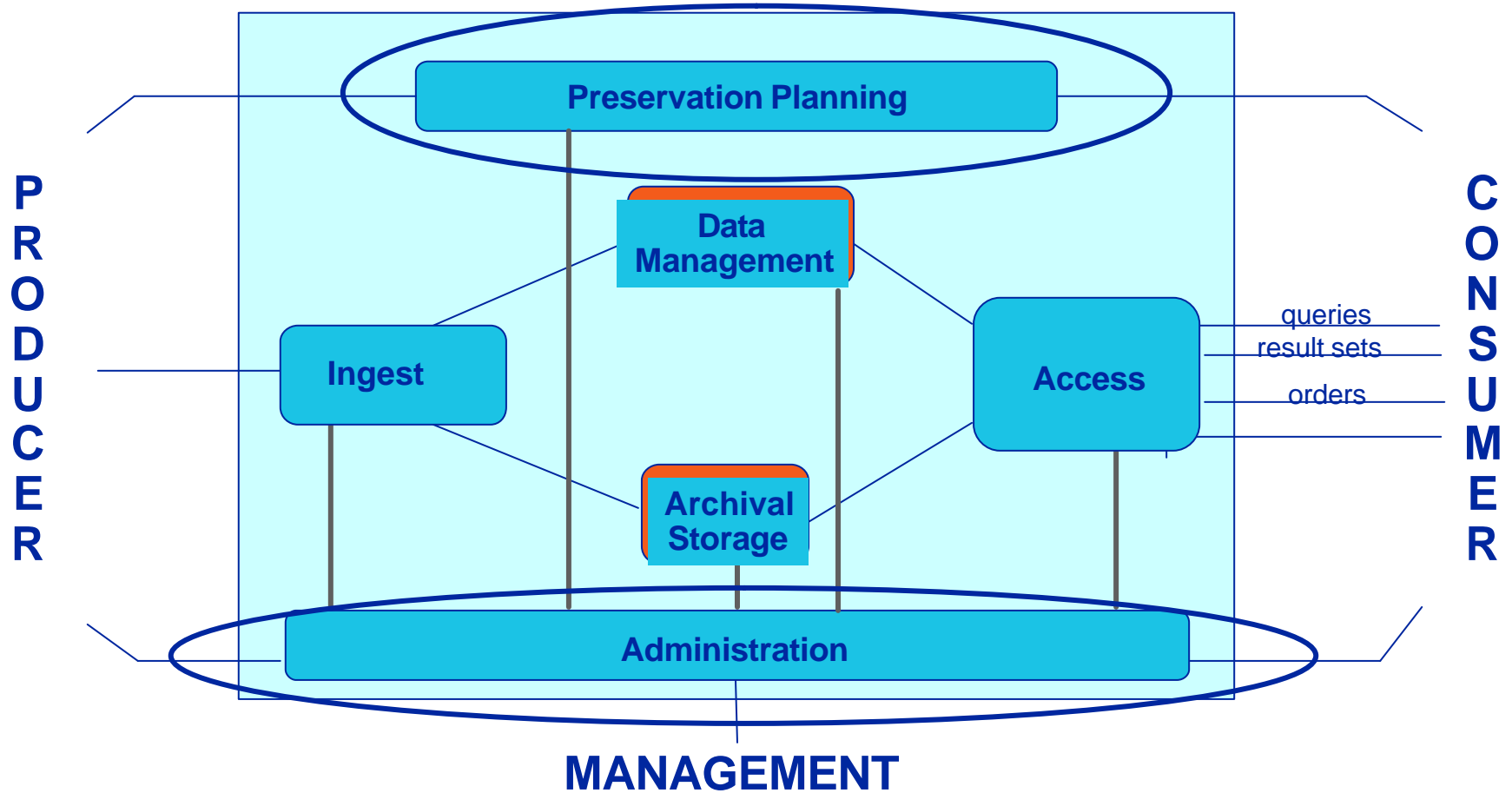
- accurately define functions and responsibilities,
- specify the external interfaces (relations with the other services and relations with the entities external to the archive),
- specify the skills required in order for the service to operate.

It is assumed that each service implements its own set of procedures and standards and has the means and resources required to carry out its tasks.

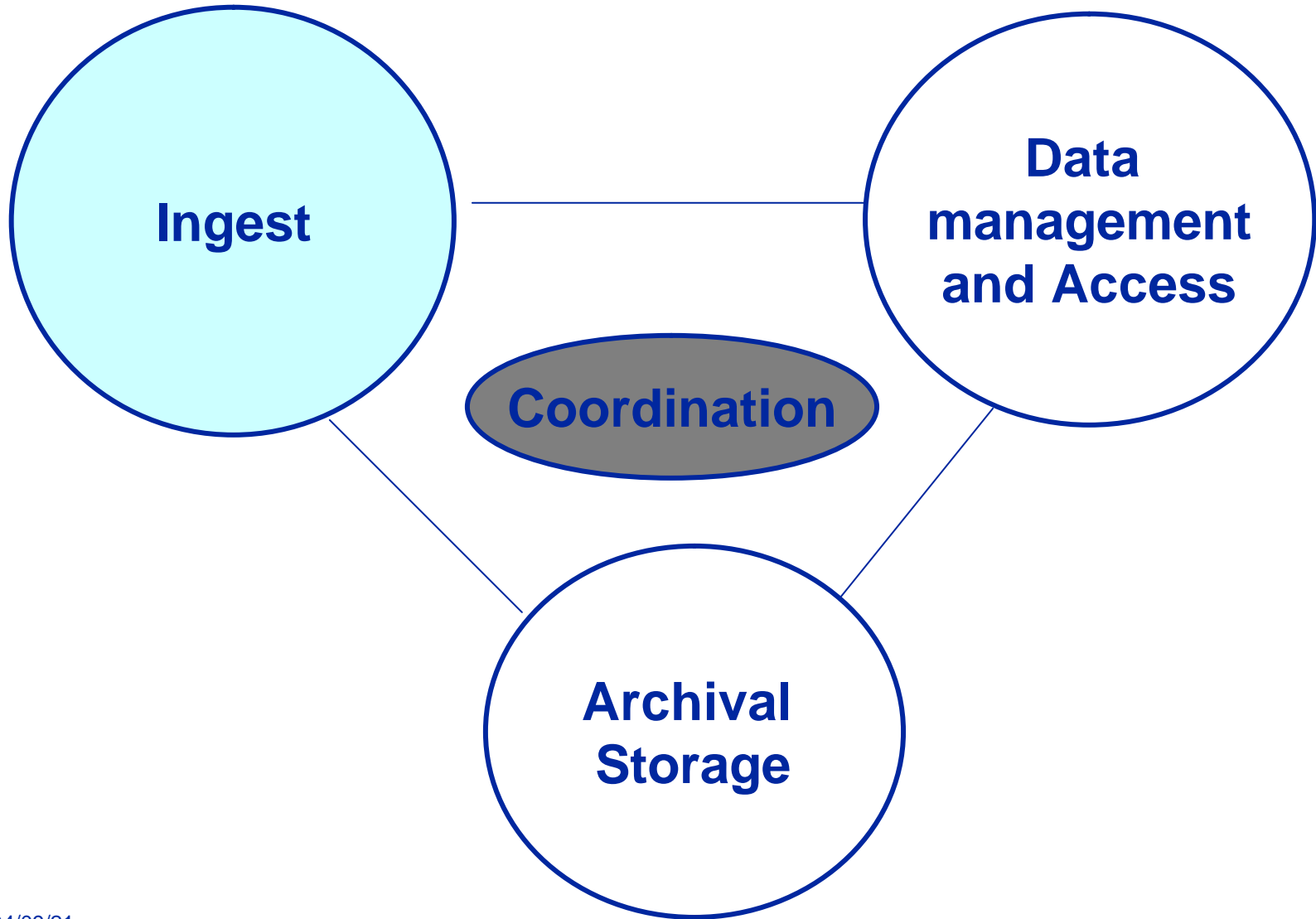
Three coordinated services



Additional considerations



Ingest service





Ingest Service

is responsible

for collecting data objects from producers and

for all tasks involved in transforming the objects handed over by producers into digital objects that meet all requirements applicable for archiving.

The Ingest Service has to deal with

- all actions and tasks identified in the CCSDS standard and the ISO draft standard 'Producer-Archive Interface Abstract Standard'
 - ❖ negotiation: what the producer can and cannot do
- transformations on the data and metadata which remain the responsibility of Archival Storage

Ingest Service

Main tasks:

- Receive objects from producers and check their compliance with the established policies
- Change the data and metadata format when necessary
 - ❖ (files delivered in MS Word format, for example, may be transformed into PDF/Archive format, and text files containing metadata may be transformed into structured XML files)
- Assign to received digital objects a unique identifier consistent with the archive's naming principles.
- Add metadata by placing received objects in a contextual relationship with other archived objects or with documents available in other archives.
- Transfer all archivable data objects (data and metadata) to Archival Storage service
- Transfer metadata and any objects that are to be accessible on line to Data Management and Access service.



Ingest Service: external interface

Interface with the Archival Storage (AS) service

- The data and metadata files to be preserved are sent to Archival Storage where they are organized in a 'virtual' tree structure.
- Archival Storage is responsible for their preservation.

Interfaces are extremely simple. They consist of a very small number of actions which can be implemented from a workstation in the Ingest service:

A realistic description of the actions used to send a digital object to AS is given below:

- ❖ *Connect to AS (always at the initiative of the Ingest Service), authentication.*
- ❖ *Request storage of a digital object, indicating its identifier and the service class expected for this object.*
- ❖ *Send file.*
- ❖ *Receive acknowledgement from AS.*
- ❖ *Close session.*

Ingest Service: external interface

Output: interface with Data Management and Access

- The metadata files, in standardized format, are sent to DMA.
- These metadata files concern all levels: they may include:
 - ❖ descriptions of collections and sub-collections (records groups and sub-groups)
 - ❖ descriptions and identification of single digital objects.
 - ❖ ...

Ingest Service: essential choices

We made the following essential choices:

- The format of digital data (office documents, scientific observations, images etc.) must be:
 - ❖ standardized, whenever possible,
 - ❖ independent of the software used to create the data,
 - ❖ described (syntax and semantics) exhaustively.
- Metadata must be standardized.

We reject methods based on format migrations conducted regularly because these migrations could not be achieved with certainty.



How many standards ...in just a few years?

hundreds and hundreds of data format standards can be found on the Web



ASCII interpretation (ISO 646)

429221140.1217.6867.9671.2815.09634.628.071205851981-09-30T14:26:17.736Z2537.991
84.430.11-80.00-80.00-14.78-17.0229221145.0617.7668.4271.7715.75629.758.361205851
31-09-30T14:26:21.975Z3902.583974.89-1.46-18.69-25.60-80.00-80.0013391147.5217.81
8.6572.0116.09627.348.511205851981-09-30T14:26:26.134Z3553.843160.87-1.35-15.79-2
01-80.00-80.0013391149.9917.8568.8872.2516.44624.938.671205851981-09-30T14:26:34
533Z3606.403068.59-1.13-15.65-21.63-80.00-80.0013391154.9217.9469.3472.7417.18620
48.991205851981-09-30T14:26:38.772Z3729.722871.54-1.10-13.96-19.98-80.00-80.0013
01157.3817.9969.5772.9817.57617.779.161205851981-09-30T14:26:42.932Z3715.963019.5
-1.10-14.96-20.79-80.00-80.0013391159.8518.0369.8073.2217.96615.409.331205851981
9-30T14:26:51.330Z4047.072315.92-0.46-11.32-13.37-80.00-80.0013391164.7718.1370.26
8.7018.79610.709.681205851981-09-30T14:26:55.570Z4073.732350.72-0.27-12.11-12.78-8
00-80.0013391167.2318.1870.4973.9319.23608.369.861205851981-09-30T14:26:59.



Separate the various information fields

58	5	1981-09-30T14:22:01.629Z	3898.70	2352.45	-1.66	-7.72	-18.61	-80.00	-80.00
58	5	1981-09-30T14:22:05.788Z	3894.64	2546.88	-1.49	-9.76	-18.96	-80.00	-80.00
58	5	1981-09-30T14:22:10.027Z	3925.05	2429.15	-1.49	-8.78	-18.27	-80.00	-80.00
58	5	1981-09-30T14:22:18.426Z	3946.38	2500.82	-1.71	-8.42	-19.33	-80.00	-80.00
58	5	1981-09-30T14:22:22.585Z	3964.01	2591.92	-1.76	-8.83	-19.86	-80.00	-80.00
58	5	1981-09-30T14:22:35.222Z	4021.69	2406.25	-1.62	-7.75	-18.24	-80.00	-80.00
58	5	1981-09-30T14:22:39.381Z	4007.80	2474.73	-1.69	-8.05	-18.88	-80.00	-80.00
58	5	1981-09-30T14:22:43.620Z	4037.11	2329.44	-1.74	-6.70	-18.23	-80.00	-80.00
58	5	1981-09-30T14:22:52.018Z	4099.12	2399.84	-1.33	-8.46	-16.84	-80.00	-80.00
58	5	1981-09-30T14:22:56.177Z	4120.44	2447.08	-1.46	-8.25	-17.44	-80.00	-80.00
58	5	1981-09-30T14:23:00.417Z	4114.44	2291.40	-1.57	-6.73	-17.11	-80.00	-80.00
58	5	1981-09-30T14:23:08.815Z	4140.36	2347.36	-1.47	-7.41	-16.91	-80.00	-80.00
58	5	1981-09-30T14:23:12.974Z	4172.42	2302.13	-1.27	-7.70	-15.82	-80.00	-80.00
58	5	1981-09-30T14:23:17.213Z	4198.63	2434.23	-1.38	-8.11	-16.80	-80.00	-80.00
58	5	1981-09-30T14:23:25.611Z	4182.57	2381.31	-1.30	-8.10	-16.30	-80.00	-80.00
58	5	1981-09-30T14:23:29.770Z	4214.61	2112.57	-1.14	-6.63	-14.20	-80.00	-80.00
58	5	1981-09-30T14:23:34.009Z	4245.87	2113.51	-0.96	-7.24	-13.35	-80.00	-80.00



The most important fields - on the semantic point of view - are named in **RED** . the other fields are named in **GREEN** .

DENSITY
LINE
ORBIT_NUMBER
SEQUENCE_NUMBER
STATION_ID
DATE
FREQUENCY
TEMPERATURE
VELOCITY
AMPLITUDES
MODE
RESONANCE_B1_POSITION
RESONANCE_B2_POSITION
GYROFREQUENCY
LOCAL_MAGNETIC_TIME
GEOMAGNETIC_LATITUDE
GEOGRAPHIC_LATITUDE
GEOGRAPHIC_LONGITUDE
ALTITUDE
VALUE_OF_L
END_OF_LINE_DATA

FREQUENCY

Definition : plasma frequency

Unit : k Hz

Type : REAL

Range : 100 .. 9675

Length : 80 bits

Format FORTRAN : F10.2

TEMPERATURE

Definition : electron temperature

Unit : K

Type : REAL

Range : 500 .. 10000

Length : 80 bits

Format FORTRAN : F10.2

VELOCITY

Definition : electron velocity

Unit : km/s

Comment : BE CAREFUL : validation in progress

Define the meaning

For each field



Ingest Service: technical resources required

The hardware, software and communication resources required to receive digital objects correctly after they have been sent by the producer services have no special characteristics.

- It is necessary to decide on a case-by-case basis whether to secure the transfers:
 - ❖ to authenticate the objects received,
 - ❖ and guarantee their fixity with respect to objects sent the producer.

These resources will need to be adapted depending on the volume of data to be handled and the frequency of transfers.

Obviously, a set of software programs will be required for computer-aided preparation of the data and metadata to be archived.

Ingest Service: skills required

There is clearly a need for two types of competence:

- The Archivist, who can:
 - ❖ define, jointly with the producer, the information to be preserved,
 - ❖ check that this information is meaningful and complete,
 - ❖ organize this information within a structured system.

- The Computer Expert, specialist in **data management** and representation of information in digital format, in order to:
 - ❖ define the data and metadata formats acceptable for preservation,
 - ❖ check their conformance,
 - ❖ implement, if necessary, a format transformation process,
 - ❖ specify the development of the computer tools required by this service, develop and use them.

These specialized skills in digital representation also imply broad computer knowledge.

Ingest Service: skills required

These two areas of competence are combined in a new profession referred to as Digital Data Manager, strongly based on norms and standards applicable to data and metadata representation.

In addition, there is a need to communicate and negotiate with data and document producers:

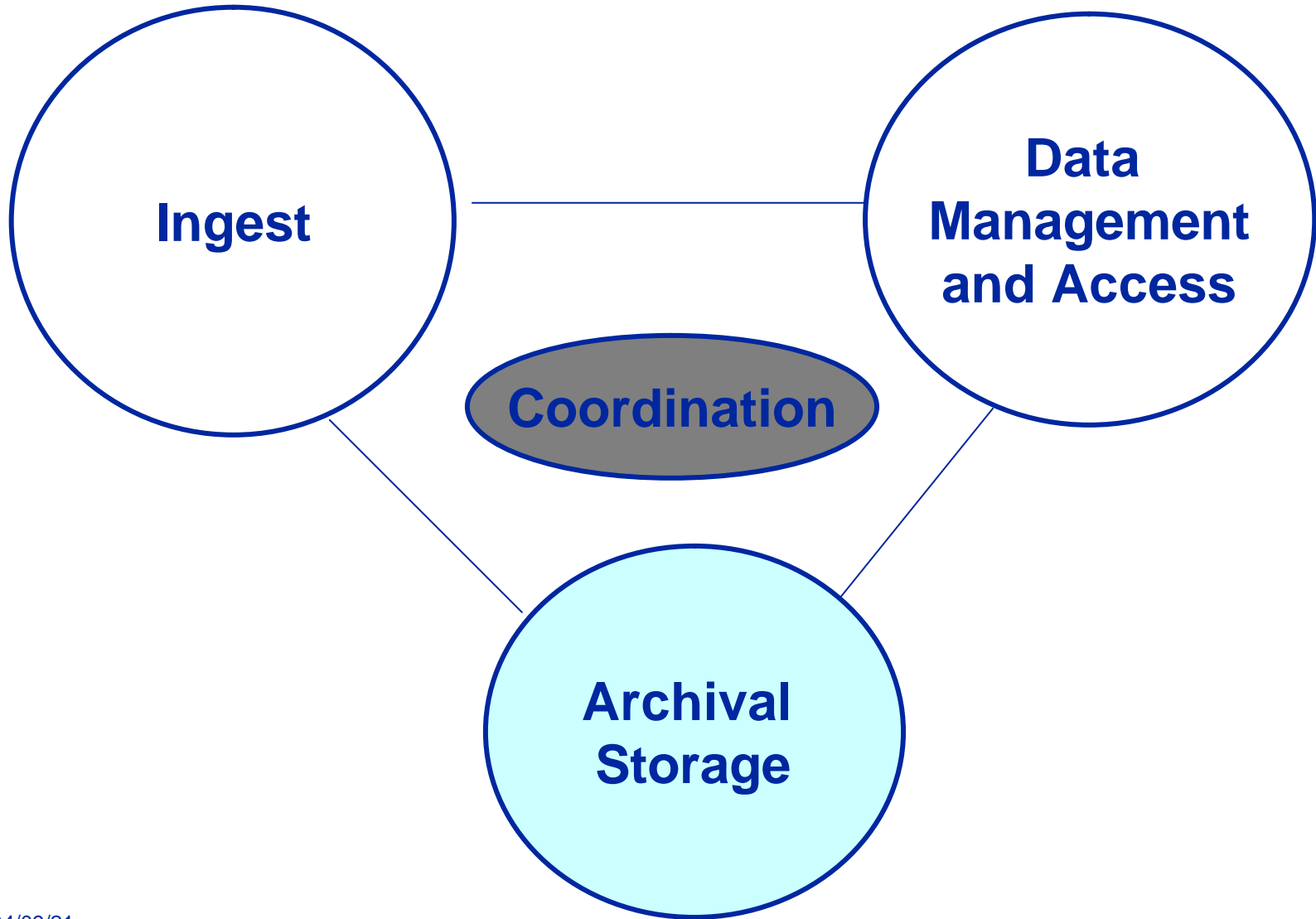
- long-term iterative work



Ingest Service: lessons learnt at CNES

In practice, the collection of a complete set of correctly described and organized digital objects represents the most difficult and, finally, the most costly task, especially in terms of human resources.

Archival Storage





Archival Storage: Functions seen from the 'customer' side

Scenario:

- I am responsible for archiving digital data, images, documents, etc.

This data consists of sets of files

- ❖ i.e. bit streams
- ❖ of known format and content
- ❖ that I can process and present to the end users in meaningful format.

- I expect an Archival Storage service to:

- ❖ take responsibility for these files for their long-term preservation
- ❖ guarantee the integrity of these files
- ❖ be able to restore these files within the delay agreed by the service contract
- ❖ provide a **stable 'technical interface'** whose services I can call upon (archive a file, restore, rename, create a virtual tree structure, etc.)
- ❖ be able to handle technology changes (migrations of storage media, etc.) with no impact on the interface and therefore no impact on my applications
- ❖ manage the access rights to this data.

- Based on this concept:

- ❖ the organization of Archival Storage remains completely independent of the other services
- ❖ the service can be reused in numerous contexts within the relevant organization.



Archival Storage: responsibility for file integrity

ES must take responsibility for all activities required to maintain the integrity of digital objects:

- storage of objects on storage media, together with one or more backup copies which must be kept in separate locations,
- continuous monitoring of the condition of the media (number of read operations carried out on each medium, measurable bit error rate, etc.),
- periodic replacement of media considered less reliable by new media,
- integration of storage technology upgrades to carry out migrations (periodic or continuous depending on the chosen policy) to the new media best adapted to its activities.
- etc.



Archival Storage: Skills

Skills of computer experts specialized in:

- the management of large sets of stored files, duplicated on various types of medium
- high-speed network technologies in order to communicate with the 'customers' of the service,
- high-capacity storage technologies, storage robots, etc., the storage media, their characteristics and reliability,
- the means used to monitor the condition of the media, implementation of these means
- the ability to maintain in operational readiness a system open round the clock and to adapt the system according to technological upgrades and increased demands
- ...



Archival Storage: the STAF service set up at CNES

STAF stands for "Service de Transfert et d'Archivage de Fichiers" (File Archive and Transfer Service)

set up in 1994

- The aim of the STAF service is to preserve CNES data files collections obtained from scientific experiments.
- This data consists of non-reproducible reference data, stable in time and intended for long term use.
- The STAF mission: to store a "collection" of files using an application logic that remains valid regardless of changes in systems and storage technologies.



Archival storage: the STAF service set up at CNES

Guarantee data integrity and confidentiality for each 'customer' using the service

Transparency of operations

Possibility of extending the storage capacities as required

Possibility of integrating new client machines

Currently: more than 3.8 million files for a volume of 145 Terabytes

Lessons learnt at CNES

The STAF (File Archive and Transfer Service) concept has stood the test of time

- Ten years of experience
- No data lost
- Increasing number of customers
- Increasing volume of stored data

• The concept also allows the archival storage system to be shared

- ❖ between several sites of the same institution
- ❖ between several separate institutions

• A minimum critical size is essential in order to reduce costs (hardware, software and human resources)

Lessons learnt at CNES

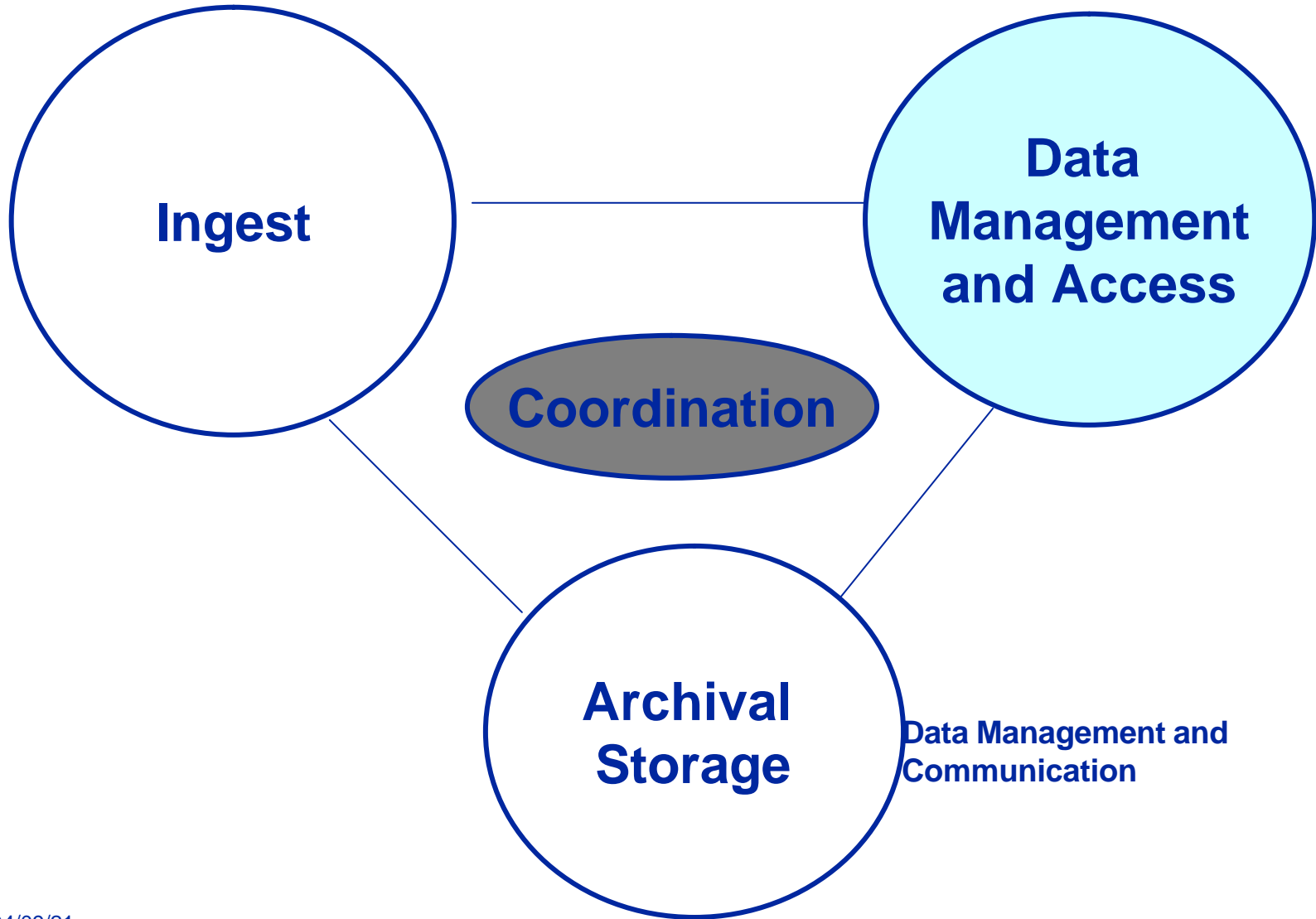
The operating principle of this service and the practical results obtained over the last 10 years have convinced the BnF of the benefits offered by this type of service.

A draft agreement is being discussed between CNES and the BnF concerning the reuse, by the BnF, of the management software implemented in the service.

This type of service may be:

- specific to an institution, or shared by several separate institutions,
- managed by a private company.

Data Management and Communication





Data Management and Access: functions

manage the data collection preserved by the Archive and
communicate this information to authorized users.

implement and maintain a computer system providing remote access
via a graphic user interface - to a set of functions

- know the content of the archive,
- find the data they need (selection criteria based on metadata, for example),
- select the data corresponding to their needs,
- order and retrieve this data,
- if necessary, transform the archived data before supplying it to the user (change format, Added-Value Services, etc.).



Data Management and Access: functions

The search for useful data is based on metadata together with additional techniques (browsing, data mining, etc.)

The data can be retrieved via the network or copied onto a standard medium (CD-ROM, DVD, DLT, etc.) depending on the volume

- a dedicated service at the CNES Computer Centre (SEM: Service d'Échange de Média / Medium Exchange Service) has been set up to allow data distribution and reception on physical media

Manage relations with the community of users



Data Management and Access: technical resources required

Technical resources required

- The system set up by DMA relies largely on database and Internet information communication technologies.
- Systems partially or totally meeting the requirements of DMA are or will be available on the market, limiting the costs of special computer developments.
- DMA may need to have access to the resources required to copy digital objects onto distribution media. Lastly, in some cases, regardless of data managed by Archival Storage, it may need to have its own storage area for data objects that must be immediately available on line for users, hence the need for a certain amount of storage capacity (usually on disk).

Data Management and Access: skills required

skills of computer experts specialized in:

- data models
- information search processes
- database technologies
- Internet technologies and languages (Human-machine Interface on browser etc.)
- maintaining in operational readiness systems open to large and small communities of users.

general knowledge of archiving issues

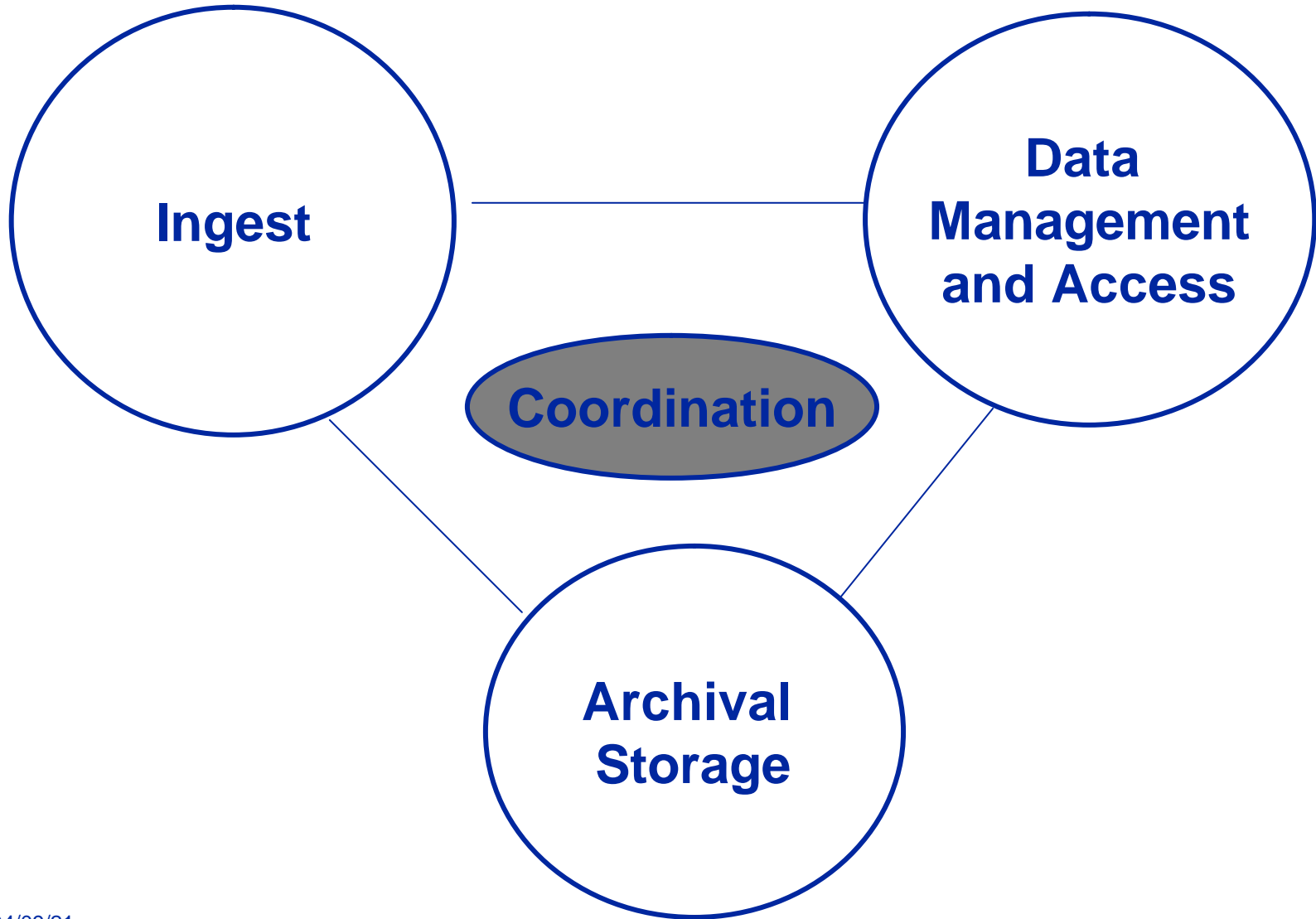
- knowledge of the data categories handled
- knowledge of metadata and data selection criteria appropriate for user needs

Data Management and Access: feedback to CNES

These Data Management and Communication skills have been implemented

- to provide space scientific data on various topics (astronomy, oceanography, etc.).
- Despite the diversity of digital objects and logics specific to each topic, the problem, which will shortly be solved, is to reduce costs through the use of a generic system that can be adapted to each topic.

Coordination



The Coordinator

The Coordinator directs and has overall responsibility for archiving (as understood in the OAIS context)

- Depending on the context, the coordinator is known as the data manager, technical data collection manager, archivist, main archivist, etc.

Mission:

- Organize the workload between the various 'services'
- Ensure that the interfaces between these services are clear
- Coordinate the work for common areas of competence:
 - ❖ the information model
 - ❖ the dictionary of deliverable digital objects
 - ❖ etc.



Why an organizational Model ?

to define the appropriate organization (human resources et technical facilities) for a digital archive,

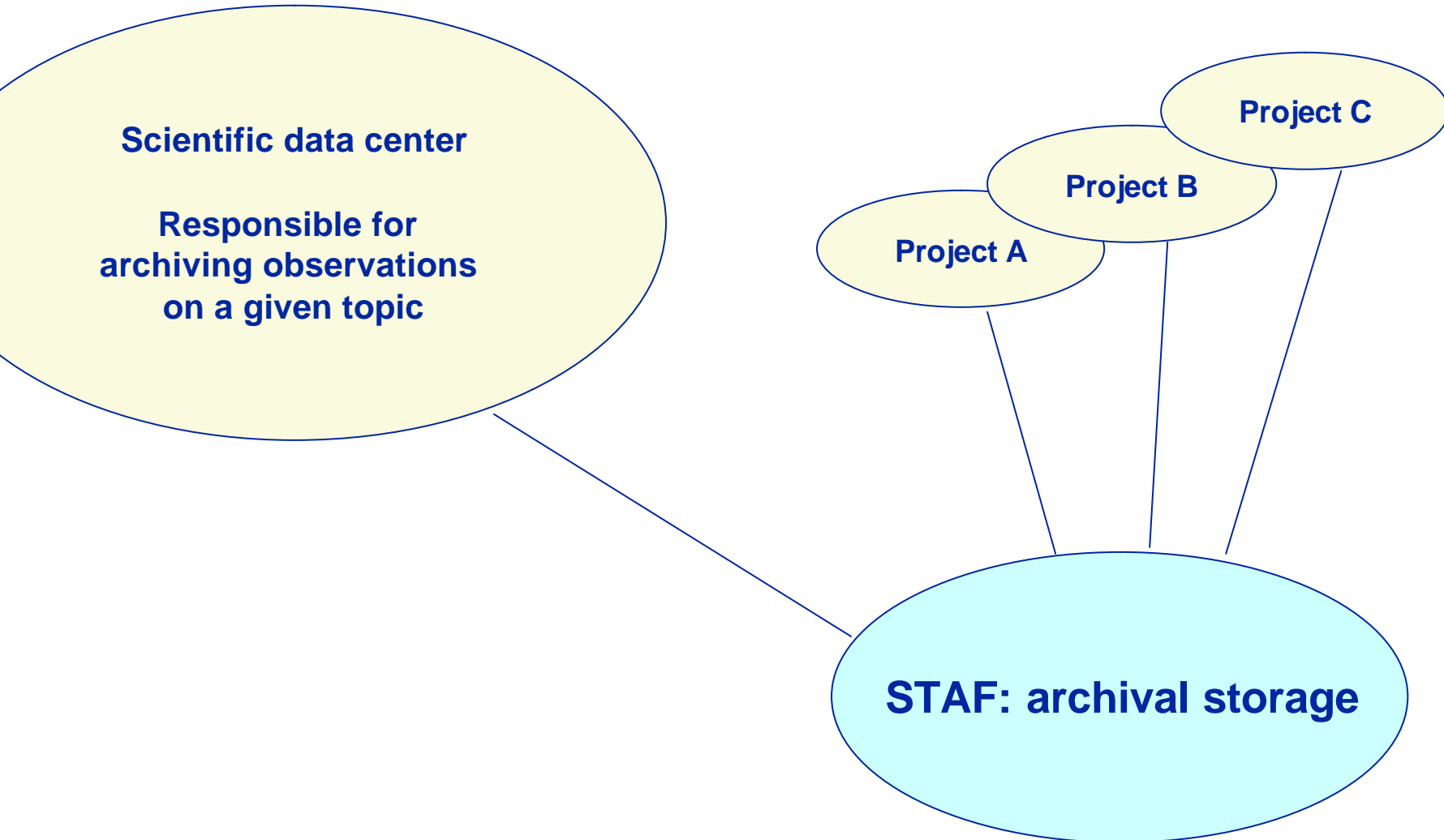
to define precisely the required skills for each service

the split in services make easier the cost analysis

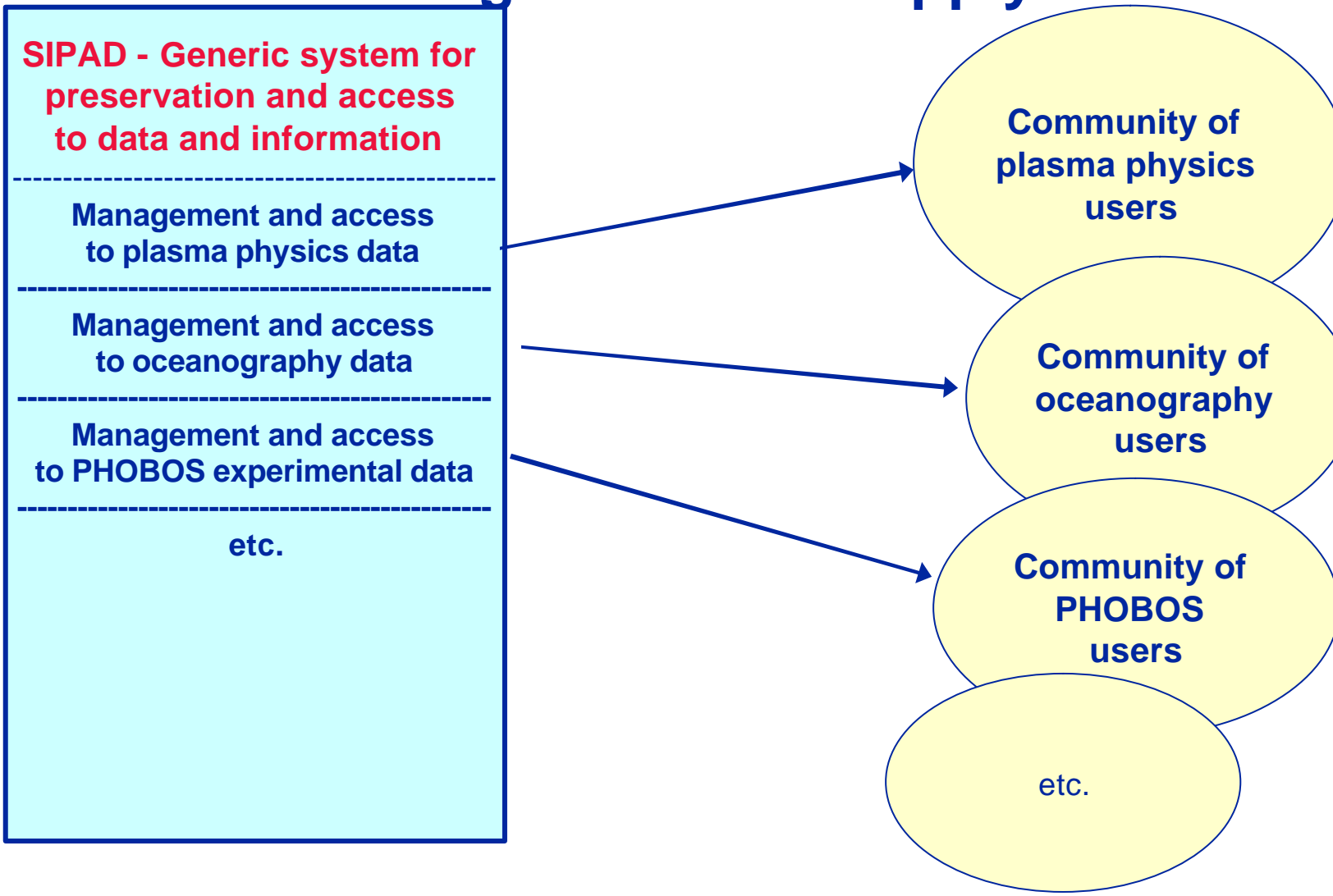
the split in services should help for the definition of commercial products well adapted to the needs of each service

to simplify the definition of an archive certification process

Relations with the institution's Information System: STAF example



Relations with the institution's IS: information management and supply





Essential Critical Point

The critical point is collecting information and ensuring all activities required to create:

- files in format suitable for long-term preservation,
- 'standardized' metadata.

This critical point concerns both:

- the content (completeness, accuracy, authenticity)
- the format (open, standardized, etc.)

The critical point depends to a certain extent on the company's technical policy and office computer policy.

Conclusion

Technology will continue to change **but information will stay**

rather than trying to keep a particular technology alive in the long term, we therefore deliberately gave priority to the option based on knowledge of the structure, syntax and semantics of the information

The proposed Organizational Model is mainly based on this choice

- Its vocation is to **help identify concrete, applicable solutions.**
- It is also based on analysis of skills and professions.
- Lastly, it is based on extensive feedback to CNES which has backed up our choice.

Conclusion

It must be possible to carry out external inspections and audits on the type of organization

The digital archive must be able to demonstrate

- through its organization,
- its resources,
- its teams and applicable standards and procedures,

its ability to perform its mission and therefore guarantee long-term preservation of the digital information it is responsible for.

This opens the door to debate on 'Certification' of Digital Archives



Conclusion: our vision for the future

The Ingest Service: there is a considerable degree of intellectual added value which cannot be replaced by automatic processes. Some software programs may make the work easier, but they cannot think for us.

The Archival Storage Service: fundamentally technological. For this type of service, companies must be able to provide turnkey solutions which are both reliable and financially attractive.

The Data Management and Access Service: highly technological, depends to a large extent on the information model. Once again, turnkey data management and access systems easily be adapted. Use in different fields can be developed.

References

[] CCSDS 650.0-B-1., Reference Model for an Open Archival Information System (OAIS). *CCSDS Blue Book. Issue 1. January 2002, ISO 14721:2003,*

[] CCSDS 651.0-R-1., Producer-Archive Interface Methodology Abstract Standard (AIMAS). *CCSDS Blue Book Issue 1, May 2004. (ISO standardization process currently started)*

[] CCSDS 647.1-B-1., *Data Entity Dictionary Specification Language (DEDSL) – Abstract Syntax (CCSD0011). CCSDS Blue Book. Issue 1. June 2001 ISO 1961:2003,*

[] CCSDS 647.3-B-1 Data Entity Dictionary Specification Language (DEDSL) – XML/DTD Syntax (CCSD0013). *CCSDS Blue Book. Issue 1. January 2002, ISO 2643:2003,*

*All these References are freely available on the web site:
<http://www.ccsds.org/CCSDS/recommandreports.html>*