Long-term Information Preservation

CCSDS Contributions

Introduction

- **We are the CCSDS Data Archive Ingest (DAI) WG**
  - Consolidates former DAI WG, Information Packaging and Registry (IPR) WG and Repository Audit and Certification (RAC) WG

- **What we've done**
  - Working within CCSDS, but is more broadly accepted than space industry
  - OAIS Reference Model
    - "now adopted as the de facto standard for building digital archives" Cyberinfrastructure Vision for 21st Century Discovery (http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf)
- **What we're doing**
  - Certification - ISO
  - Guidance on information to capture to aid preservation
- **What we plan to do**
  - Continued development of digital archiving standards
  - We'd like you to be an advocate for our work.
  - We'd appreciate inputs from you on suggested directions.

- **Details** What we're doing and why we approach digital preservation the way we do

Reference Model for an Open Archival Information System (OAIS)

- OAIS reference model provides: fundamental concepts for preservation
  - fundamental definitions so people can speak without confusion
    - Provides vocabulary – widely applicable
    - defines important **roles** in digital preservation
  - Conformance to OAIS Standard defined
  - OAIS approach to digital preservation:
    - covers all types of digitally encoded information
    - provides a way to test whether preservation is successful
    - does not require seeing into the future
    - does require transparency
    - but does not require “open access”
  - does not cover
    - social and organizational aspects
    - finance etc.
    - many detailed aspects noted in the "roadmap of follow-on standards"
- is not meant as a **design/blueprint** for a repository
  - it allows greater flexibility for implementations
- OAIS does provide a good basis for certification

Challenges

- Wanted it to be applicable to all kinds of
  - Digital objects
  - Organizations
- Hardware
- Software
- Sources
- Users
- Legal systems
- All kinds of changes

Key OAIS Concepts

- Claiming “This is being preserved” is untestable
  - Essentially meaningless
    - Except “BIT PRESERVATION”
- How can we make it testable?
  - Claim to be able to continue to “do something” with it
    - Understand/use
      - Need extra information to help this (Representation Information)
- Still meaningless…
  - Things are too interrelated
    - Representation Information potentially unlimited
  - Need to define a target group (Designated Community) – those we guarantee can understand – so we can test
- Many other concepts identified
- Finer grained taxonomy than simply saying “metadata”
  - Allows one to ask if one has all the required types

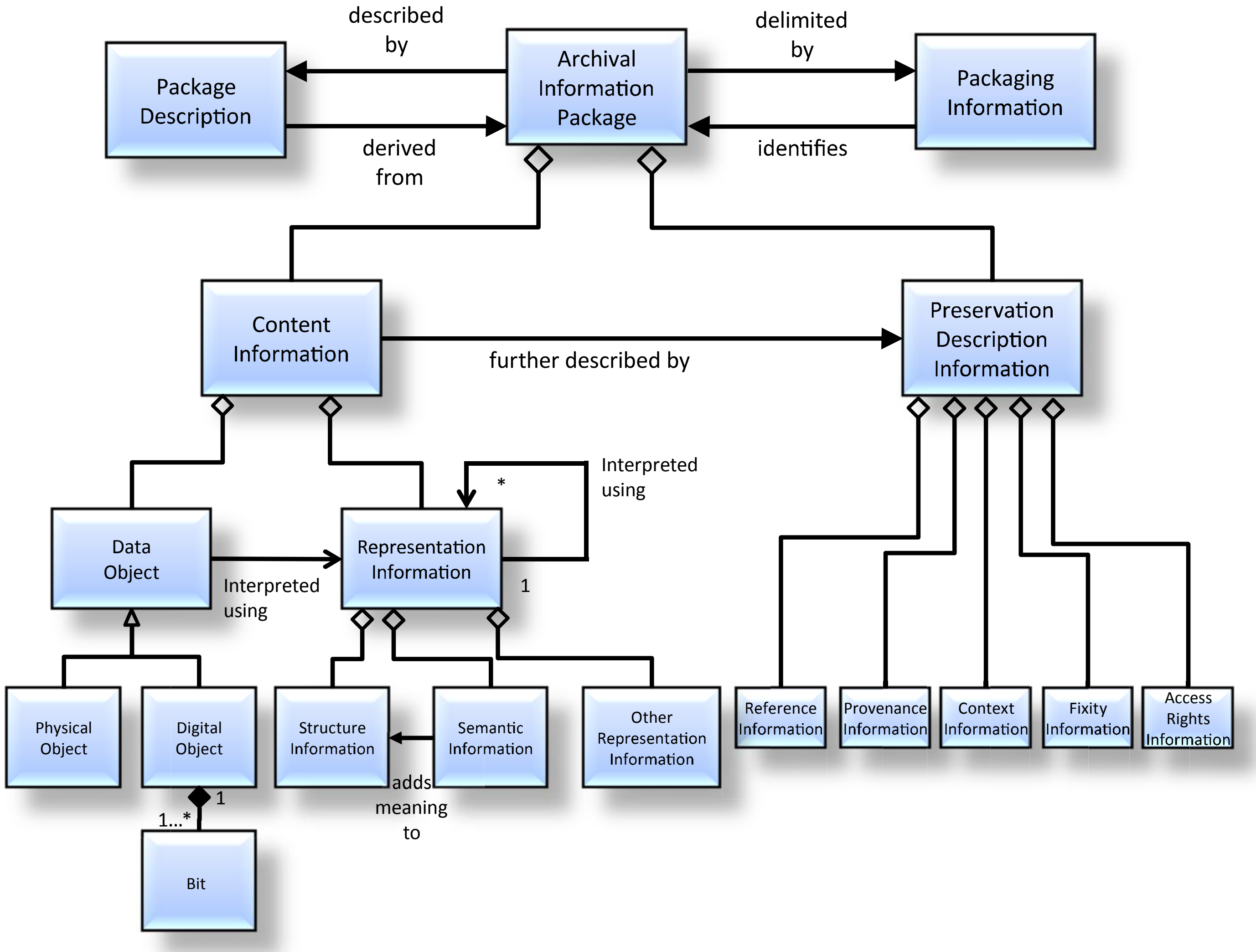
OAIS Standard Overview

- Provides information model
- Provides mandatory requirements
- Provides framework for additional standards

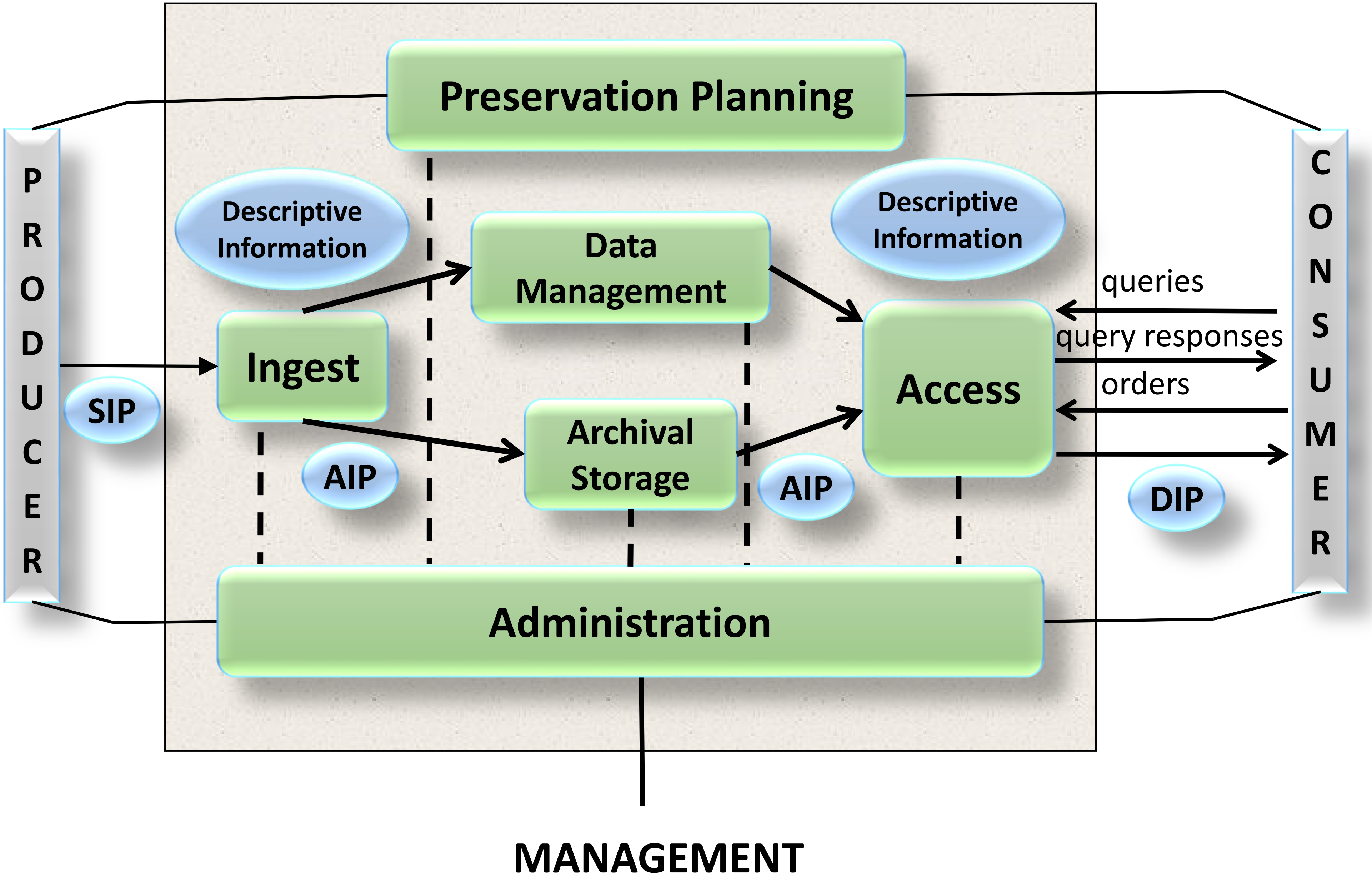
- Follow-ons
  - Auditing and Certification Standards
  - Detailed Processes
  - Some Protocols

- Builds on previous data description standards

Archival Information Package



- Logical package e.g. using pointers
- Contains everything needed for long term preservation
  - Usability
  - Authenticity

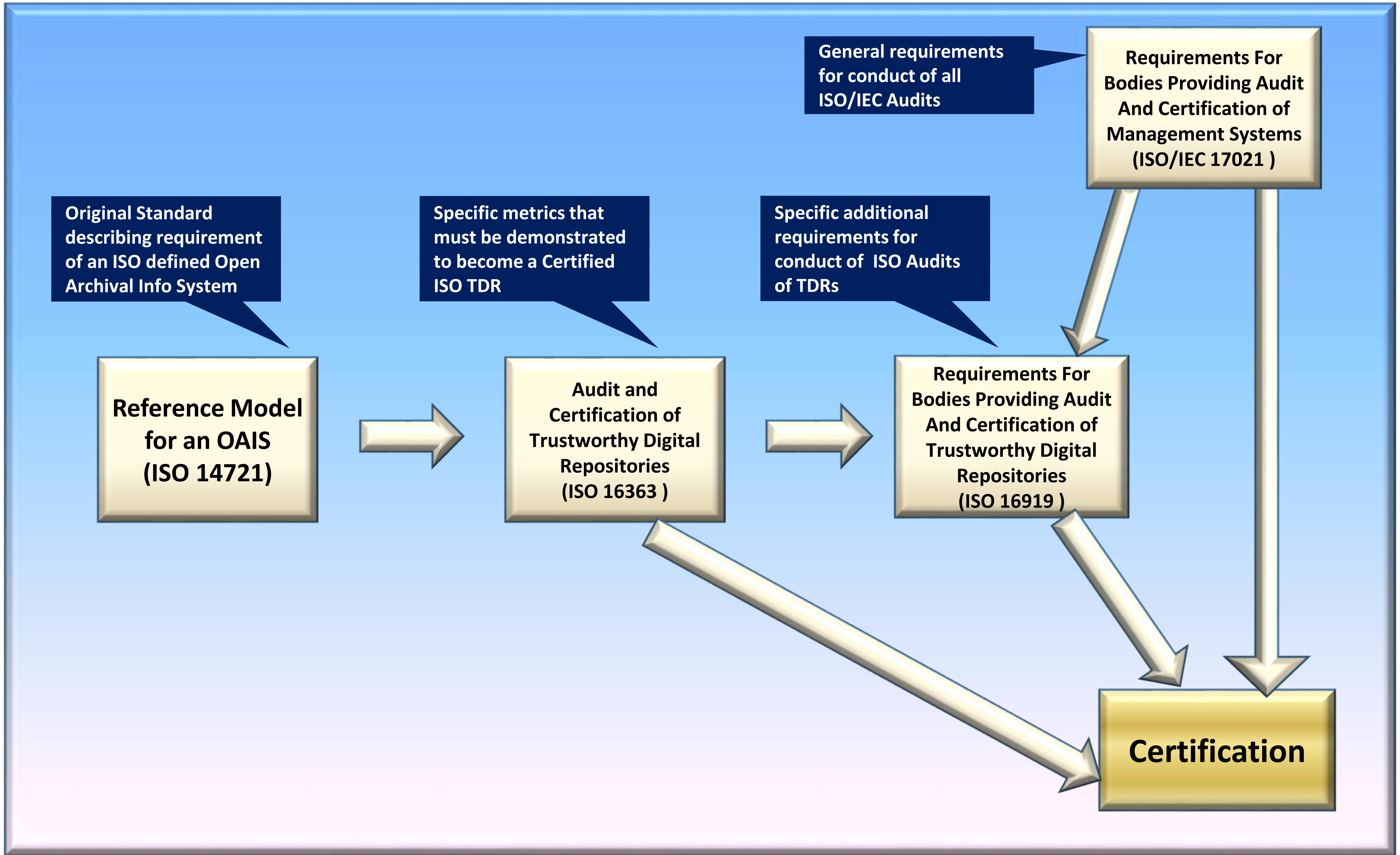

Preservation Planning
PRODUCER
CONSUMER
Descriptive Information
Data Management
Descriptive Information
queries
Ingest
query responses
orders
Access
SIP
Archival Storage
AIP
AIP
DIP
Administration
MANAGEMENT

Menu

- What we’ve done – the standards
- Relationship to previous work – origins of digital preservation
- Examples of archives using OAIS
- Data – more details about the issues
- Preservation techniques
- New standards in development
- Advocating

What we’ve done – the standards

Standards based Repository Audit and Certification (ISO 16363 and ISO 16919)



General requirements for conduct of all ISO/IEC Audits

Requirements For Bodies Providing Audit And Certification of Management Systems (ISO/IEC 17021)

Original Standard describing requirement of an ISO defined Open Archival Info System

Specific metrics that must be demonstrated to become a Certified ISO TDR

Specific additional requirements for conduct of ISO Audits of TDRs

Reference Model for an OAIS (ISO 14721)

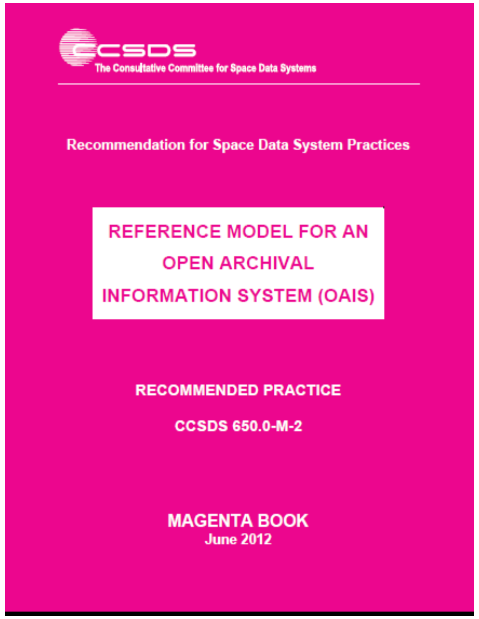
Audit and Certification of Trustworthy Digital Repositories (ISO 16363)

Requirements For Bodies Providing Audit And Certification of Trustworthy Digital Repositories (ISO 16919)

Certification

Get Your Standards

- You can get all ISO Standards from ISO website at: http://www.iso.org/home/store/catalogue_tc/
- You can also download the CCSDS Magenta Book equivalents of the ISO docs for free
  - ISO 14721 (OAIS) equivalent here: http://public.ccsds.org/publications/archive/650x0m2.pdf
  - ISO 16363 (RAC metrics) equivalent here: http://public.ccsds.org/publications/archive/652x0m1.pdf
  - ISO 16919 (RAC process) equivalent here: http://public.ccsds.org/publications/archive/652x1m2.pdf



CCSDS
The Consultative Committee for Space Data Systems

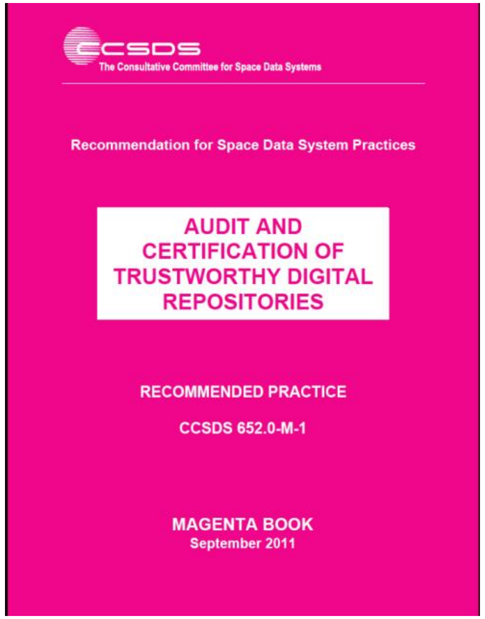
Recommendation for Space Data System Practices

REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS)

RECOMMENDED PRACTICE

CCSDS 650.0-M-2

MAGENTA BOOK
June 2012



CCSDS
The Consultative Committee for Space Data Systems

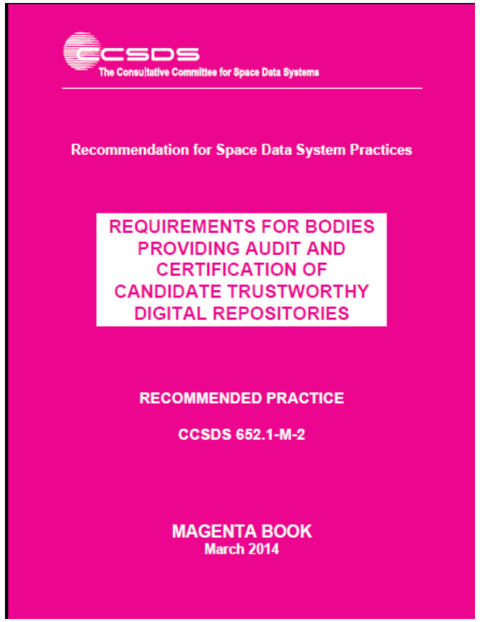
Recommendation for Space Data System Practices

AUDIT AND CERTIFICATION OF TRUSTWORTHY DIGITAL REPOSITORIES

RECOMMENDED PRACTICE

CCSDS 652.0-M-1

MAGENTA BOOK
September 2011



CCSDS
The Consultative Committee for Space Data Systems

Recommendation for Space Data System Practices

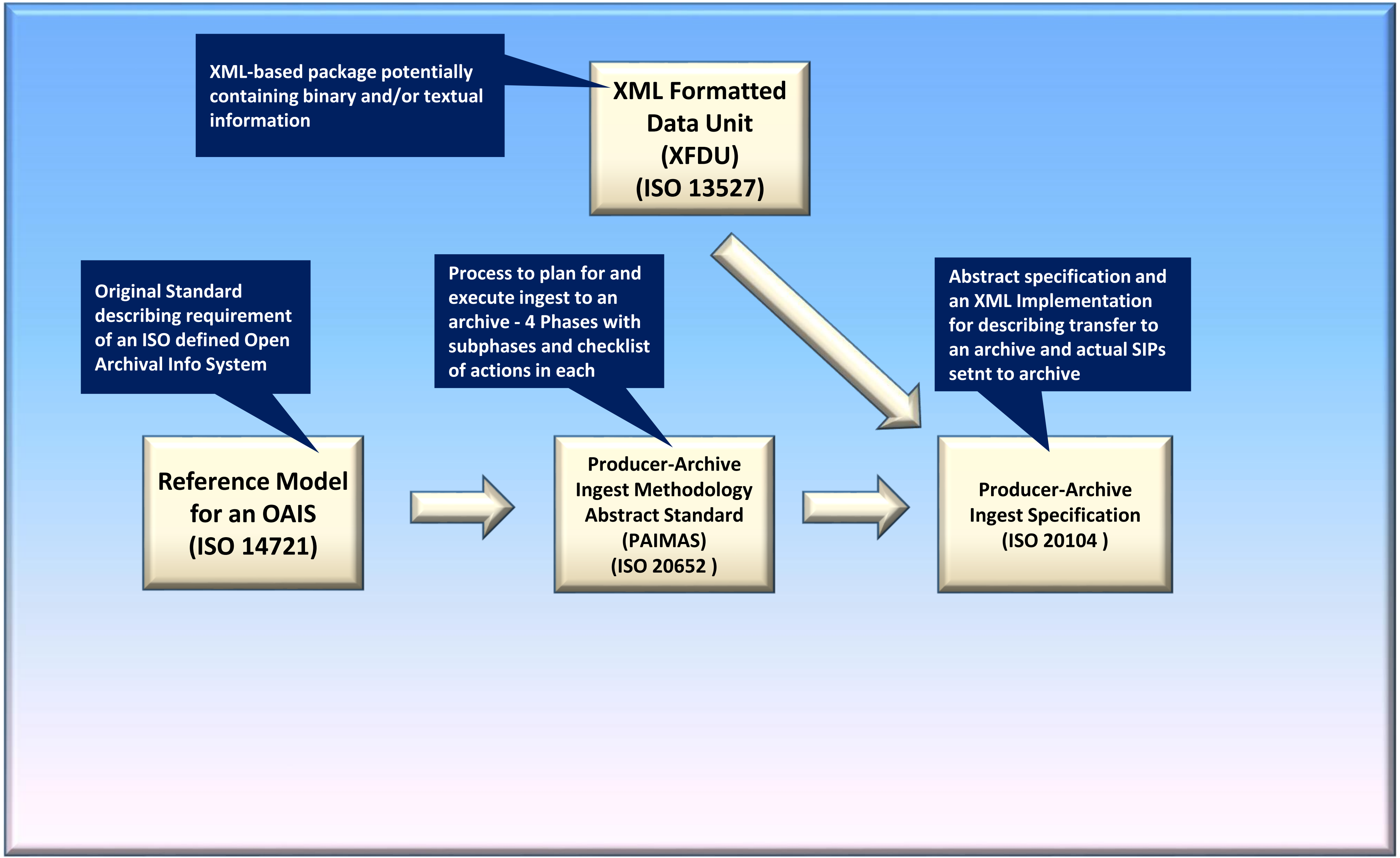
REQUIREMENTS FOR BODIES PROVIDING AUDIT AND CERTIFICATION OF CANDIDATE TRUSTWORTHY DIGITAL REPOSITORIES

RECOMMENDED PRACTICE

CCSDS 652.1-M-2

MAGENTA BOOK
March 2014

Archive Process and Protocol Standards (ISO 20652 and ISO 20104)



XML-based package potentially containing binary and/or textual information

XML Formatted Data Unit (XFDU) (ISO 13527)

Original Standard describing requirement of an ISO defined Open Archival Info System

Process to plan for and execute ingest to an archive - 4 Phases with subphases and checklist of actions in each

Abstract specification and an XML Implementation for describing transfer to an archive and actual SIPs setnt to archive

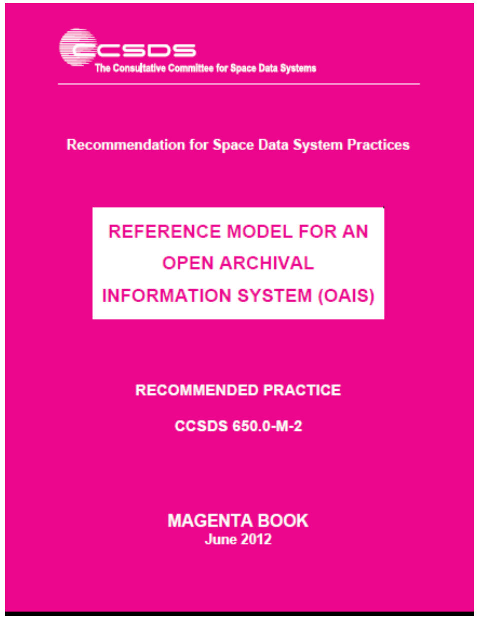
Reference Model for an OAIS (ISO 14721)

Producer-Archive Ingest Methodology Abstract Standard (PAIMAS) (ISO 20652)

Producer-Archive Ingest Specification (ISO 20104)

Get Your Standards

- You can get all ISO Standards from ISO website at: http://www.iso.org/home/store/catalogue_tc/
- You can also download the CCSDS Book equivalents of the ISO docs for free
  - ISO 14721 (OAIS) equivalent here: http://public.ccsds.org/publications/archive/650x0m2.pdf
  - ISO 20652 (PAIMAS) equivalent here: http://public.ccsds.org/publications/archive/651x0m1.pdf
  - ISO 20104 (PAIS) equivalent here: http://public.ccsds.org/publications/archive/651x1b1.pdf
  - ISO 13527 (XFDU) equivalent here: http://public.ccsds.org/publications/archive/661x1b1.pdf



CCSDS
The Consultative Committee for Space Data Systems

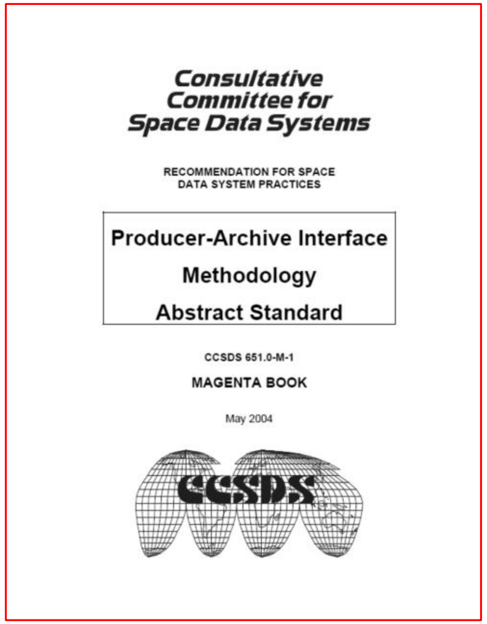
Recommendation for Space Data System Practices

REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS)

RECOMMENDED PRACTICE

CCSDS 650.0-M-2

MAGENTA BOOK
June 2012



Consultative Committee for Space Data Systems

RECOMMENDATION FOR SPACE DATA SYSTEM PRACTICES

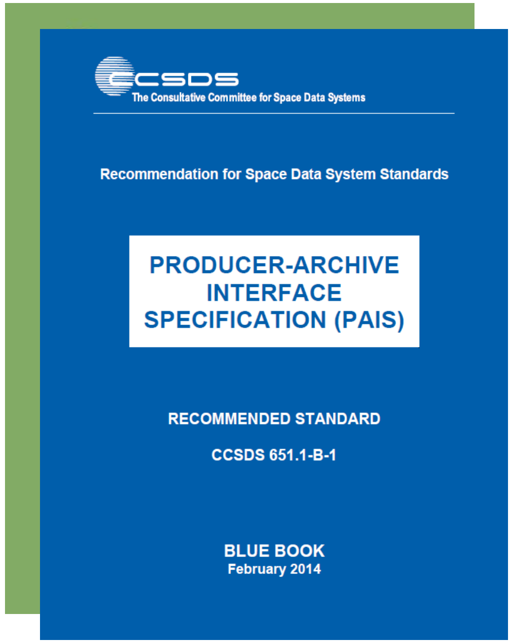
Producer-Archive Interface Methodology Abstract Standard

CCSDS 651.0-M-1

MAGENTA BOOK

May 2004

CCSDS



CCSDS
The Consultative Committee for Space Data Systems

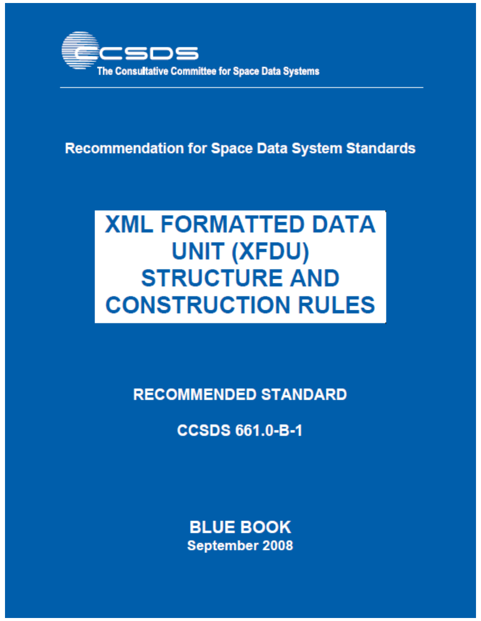
Recommendation for Space Data System Standards

PRODUCER-ARCHIVE INTERFACE SPECIFICATION (PAIS)

RECOMMENDED STANDARD

CCSDS 651.1-B-1

BLUE BOOK
February 2014



CCSDS
The Consultative Committee for Space Data Systems

Recommendation for Space Data System Standards

XML FORMATTED DATA UNIT (XFDU) STRUCTURE AND CONSTRUCTION RULES

RECOMMENDED STANDARD

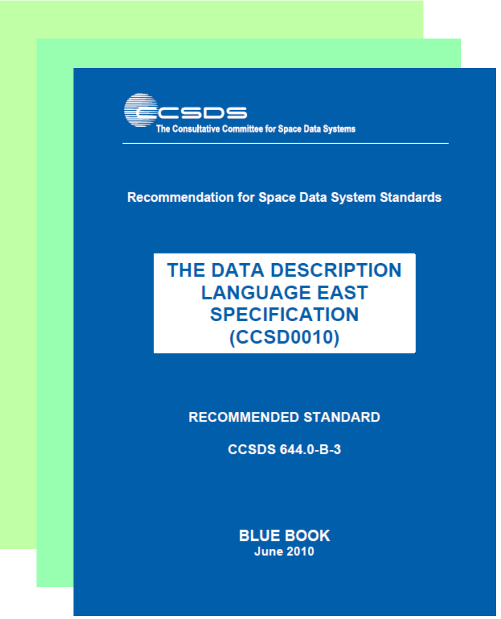
CCSDS 661.0-B-1

BLUE BOOK
September 2008

Other CCSDS DAI WG Standards

- EAST (Data Format Structure Description Language)

- Data Entity Dictionary Specification Language (DEDSL)
  - Abstract definition
  - Parameter Value Language (PVL), XML/DTD, and XML Schema implementation



CCSDS The Consultative Committee for Space Data Systems

Recommendation for Space Data System Standards

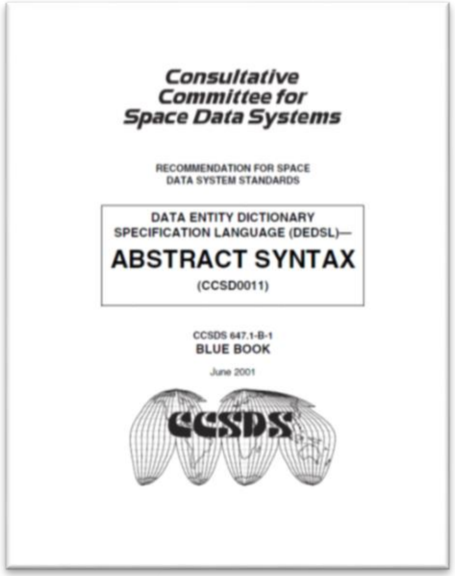
THE DATA DESCRIPTION LANGUAGE EAST SPECIFICATION (CCSD0010)

RECOMMENDED STANDARD

CCSDS 644.0-B-3

BLUE BOOK

June 2010



Consultative Committee for Space Data Systems

RECOMMENDATION FOR SPACE DATA SYSTEM STANDARDS

DATA ENTITY DICTIONARY SPECIFICATION LANGUAGE (DEDSL)—

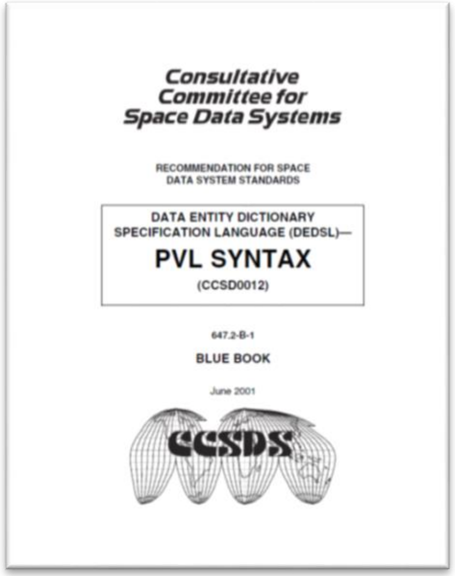
ABSTRACT SYNTAX

(CCSD0011)

CCSDS 647.1-B-1

BLUE BOOK

June 2001



Consultative Committee for Space Data Systems

RECOMMENDATION FOR SPACE DATA SYSTEM STANDARDS

DATA ENTITY DICTIONARY SPECIFICATION LANGUAGE (DEDSL)—

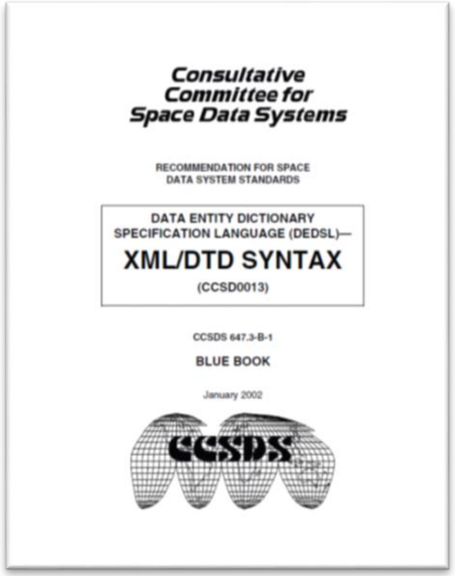
PVL SYNTAX

(CCSD0012)

647.2-B-1

BLUE BOOK

June 2001



Consultative Committee for Space Data Systems

RECOMMENDATION FOR SPACE DATA SYSTEM STANDARDS

DATA ENTITY DICTIONARY SPECIFICATION LANGUAGE (DEDSL)—

XML/DTD SYNTAX

(CCSD0013)

CCSDS 647.3-B-1

BLUE BOOK

January 2002

Relationship to previous work – origins of digital preservation

Who Begat Whom? Certification Standards

- Bruce Ambacher

Examples of archives using OAIS

Influence of OAIS-RM and Certification Audit on CIESIN Archive

- Bob Downs

Influence of the OAIS-RM on PDS4

- The Planetary Data System (PDS4) is NASA's official archive for all Solar System Exploration science data.

- PDS4 is a major revision and transition to a modern system based on best practices for data system development while leveraging 20 years of lessons learned.
    - PDS4 has been operational since 2013.
    - PDS4 has been adopted by the international Planetary Science Community and considered to be a first-of-its-kind where a single system and set of standards has been adopted by an entire science community.

- The OAIS-RM heavily influenced the architecture and design of PDS4.
    - PDS4 complies with the *Mandatory Responsibilities* required to operate an OAIS archive.
    - The PDS4 system and services architecture maps to the *OAIS Functional Model*.
    - The PDS4 Information Model is based on OAIS-RM concepts.
        - The *Information Object*, consisting of a *Data Object* and its *Representation Information*, is a foundational concept.
        - *Preservation Description Information* (PDI) is required for all data.

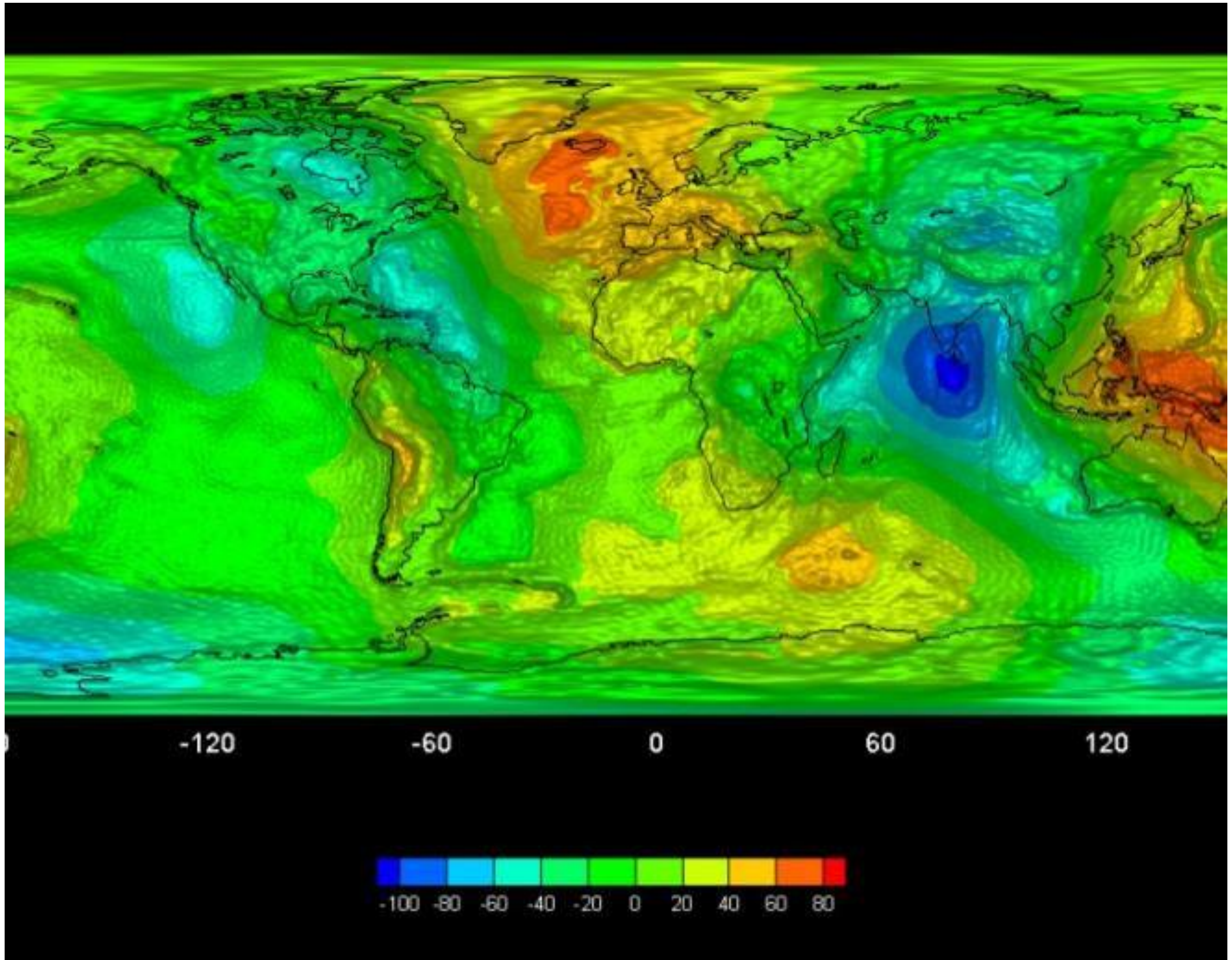
- The PDS4 functional infrastructure is designed to respond to the OAIS-RM based PDS4 Information Model.
    - The model is independent from the implementation technology allowing both to evolve separately.
    - Model extracts are used to design and configure system software and services.

- The PDS is well positioned to be accredited against the ISO-16363 standards.

Data – more details about the issues

Data contains numbers, etc.

- May be digitally encoded
  - But a lot of additional information is needed to understand it
    - Layout, Bit Format, Structure
    - Units, Field Names, Meaning, Semantics
    - Etc.



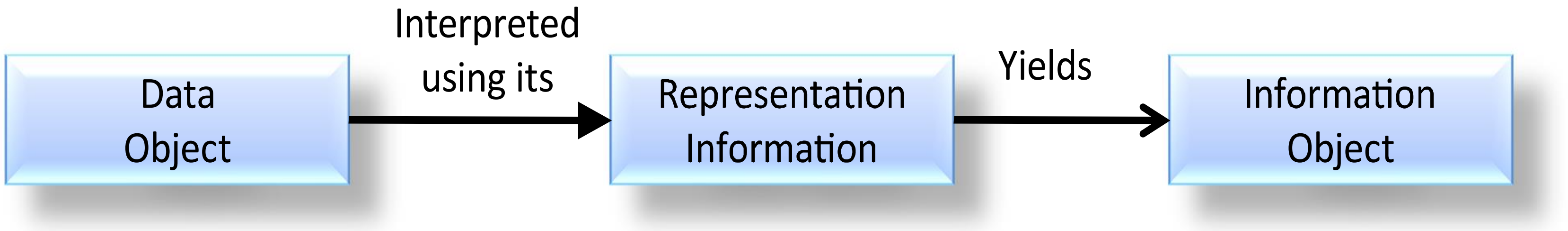
To be combined and processed to get this

Preservation techniques

Options

- **EITHER**
  - Hand on the information to some other custodian
- **OR**
  - Keep the information – in which case one can
    - **EITHER**
      - Keep the bits unchanged
        - In which case one must add what OAIS calls "Representation Information" to ensure the information encoded in the bits can still be understood/used (by the Designated Community)
    - **OR**
      - Change the bits
        - In which case one must keep evidence about why this new object should be considered "authentic"
        - And enable the new bit sequences to be understandable/usable (add appropriate "Representation Information")

Information



Digital Preservation Methods

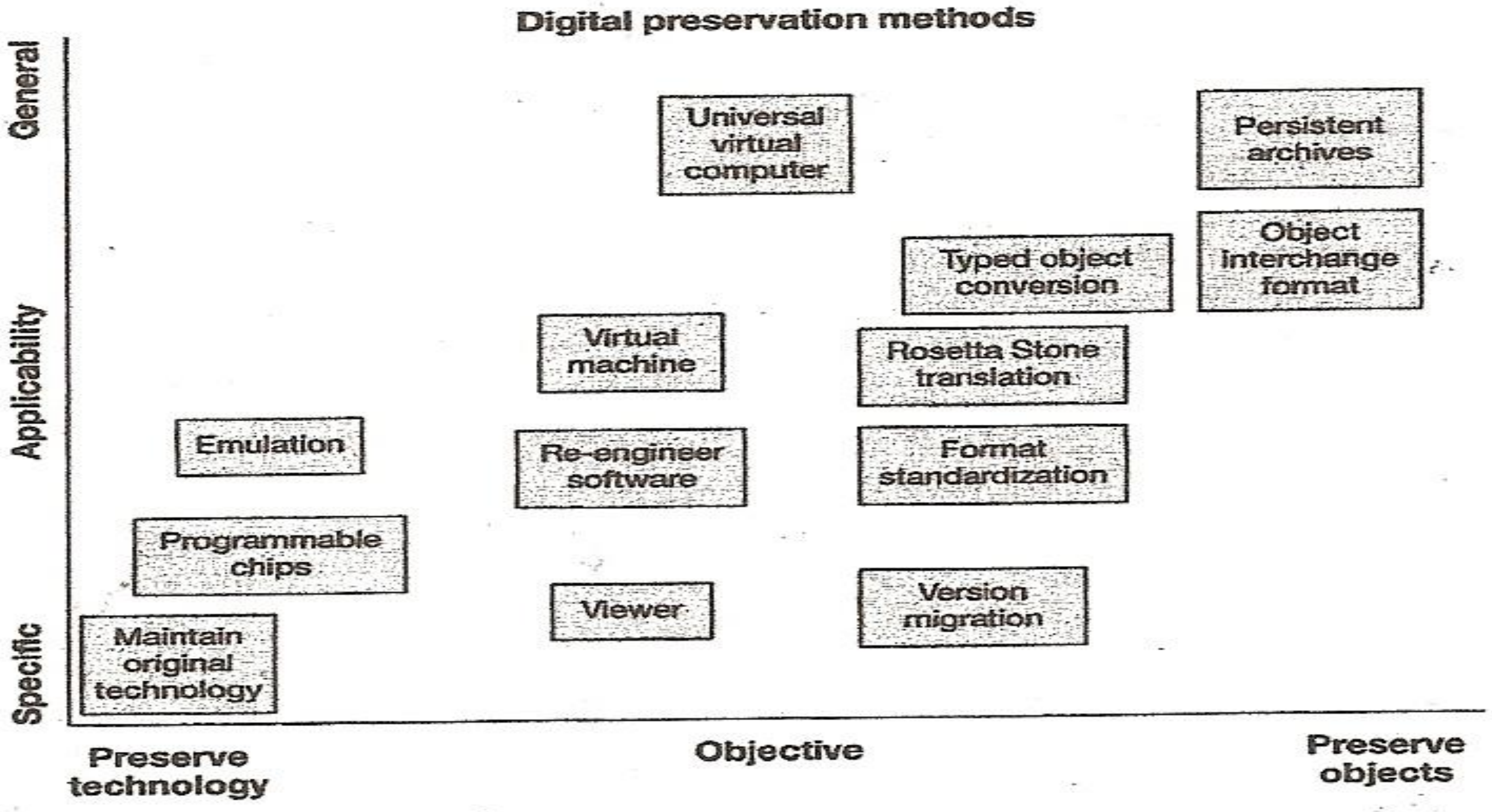


Figure 6.3 Digital Preservation Methods (From Thibodeau, 2002, p.19)

Representation Information examples

- Emulation support
- Documentation
- Data Dictionaries
- Data descriptions

Emulation – advantages and drawbacks

- Advantage
    - Can do what was do-able previously e.g.
        - display document in an format which is no longer supported by using emulation to enable use of old software. Might need to emulate chip instruction sets, operating systems etc.
        - Analyse old data using old analysis software e.g. CERN/LHC analysis software to repeat analyses of data, perhaps tweaking parameters etc
- Disadvantage
    - Can ONLY do what was do-able previously
    - BUT
        - Difficult to use OLD data in <u>**NEW**</u> analysis software
        - Difficult to combine OLD data with NEW data

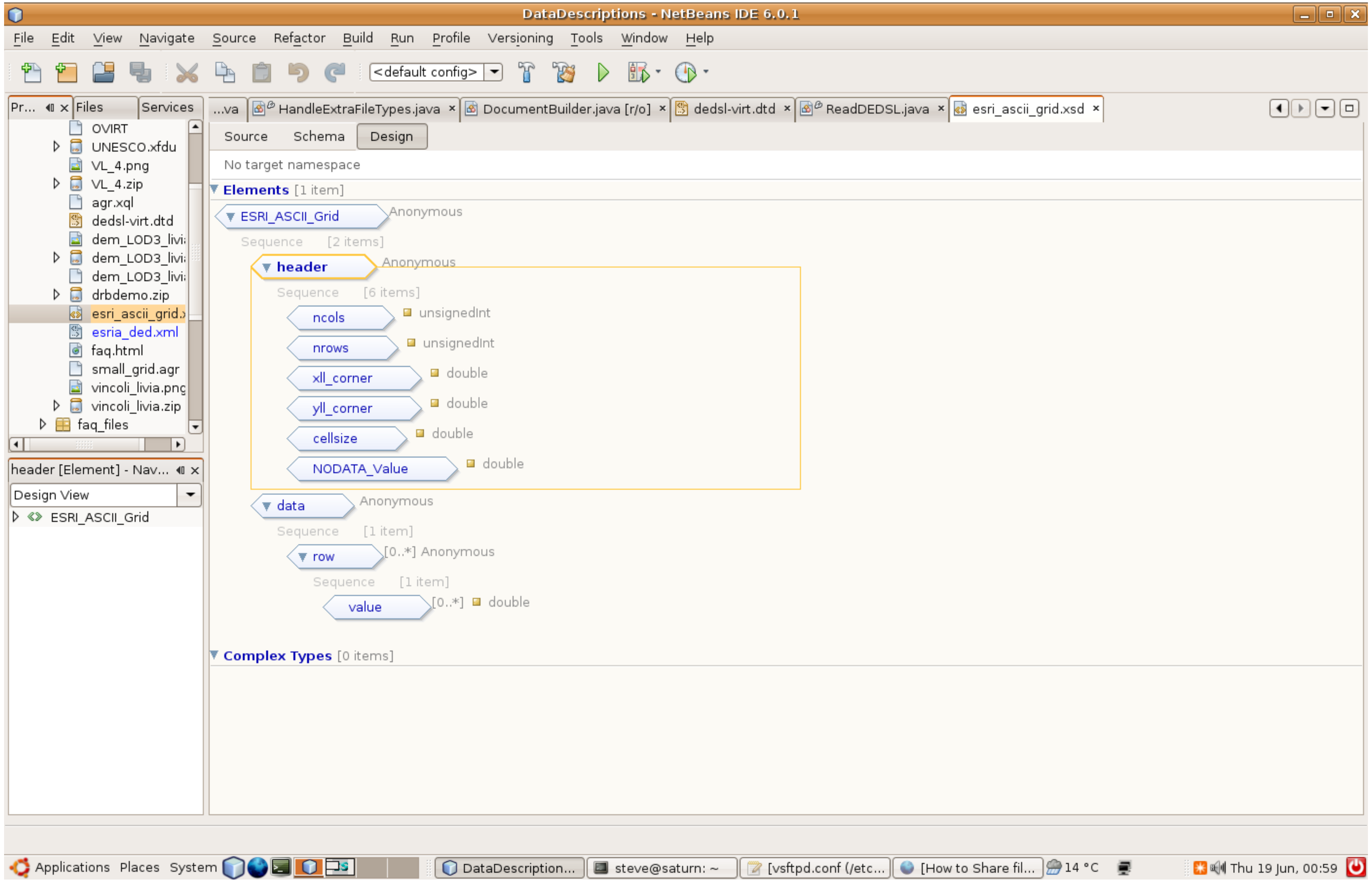
Formal Descriptions of Structure and Semantics

- CNES EAST tools (http://east.cnes.f), OASIS, EAST C Library (reference implementation).
- Also DEBAT (BEST Tools) http://debat.c-s.fr/
- Data Request Broker (DRB) - http://www.gael.fr/drb/site/
- JNI Wrapper for EAST C Library (jnieast).
- DEDSL Abstract, PVL, and XML(DTD) syntax for defining some simple data semantics. RDF, RDFS and OWL.
- Interfaces for a more general data description language and semantics API (DSSIL).
- GUI Tools for capturing Object Oriented Semantics (RDF and RDFS) and Code Generation.
- DFDL (Data Format Description Language from Open Grid Forum) https://www.ogf.org/ogf/doku.php/standards/dfdl/dfdl

Advantages of Formal Descriptions of Structure

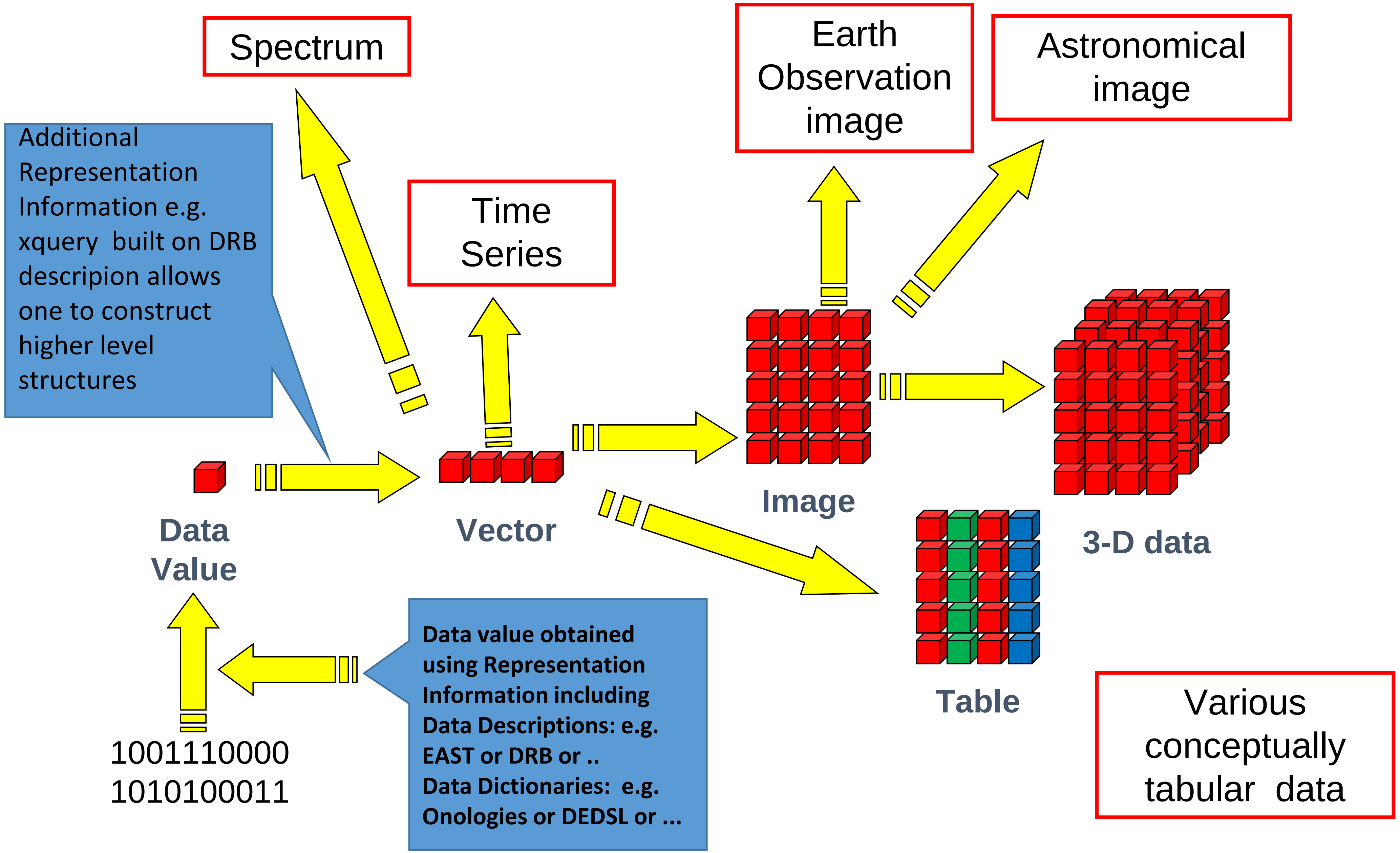
- Formal descriptions of structure gives a user the knowledge an ability to map data bits to data values in software.
- Gives a common language for syntactic information to describe many different data formats.
- One software API for many data formats:
    - may increase likelihood of data reuse
- Machine readable descriptions.

Formal Descriptions of Structure (Examples)



```
No target namespace
Elements [1 item]
  ESRI_ASCII_Grid   Anonymous
    Sequence   [2 items]
      header   Anonymous
        Sequence   [6 items]
          ncols          unsignedInt
          nrows          unsignedInt
          xll_corner     double
          yll_corner     double
          cellsize       double
          NODATA_Value   double
      data   Anonymous
        Sequence   [1 item]
          row   [0..*] Anonymous
            Sequence   [1 item]
              value   [0..*]   double
Complex Types [0 items]
```

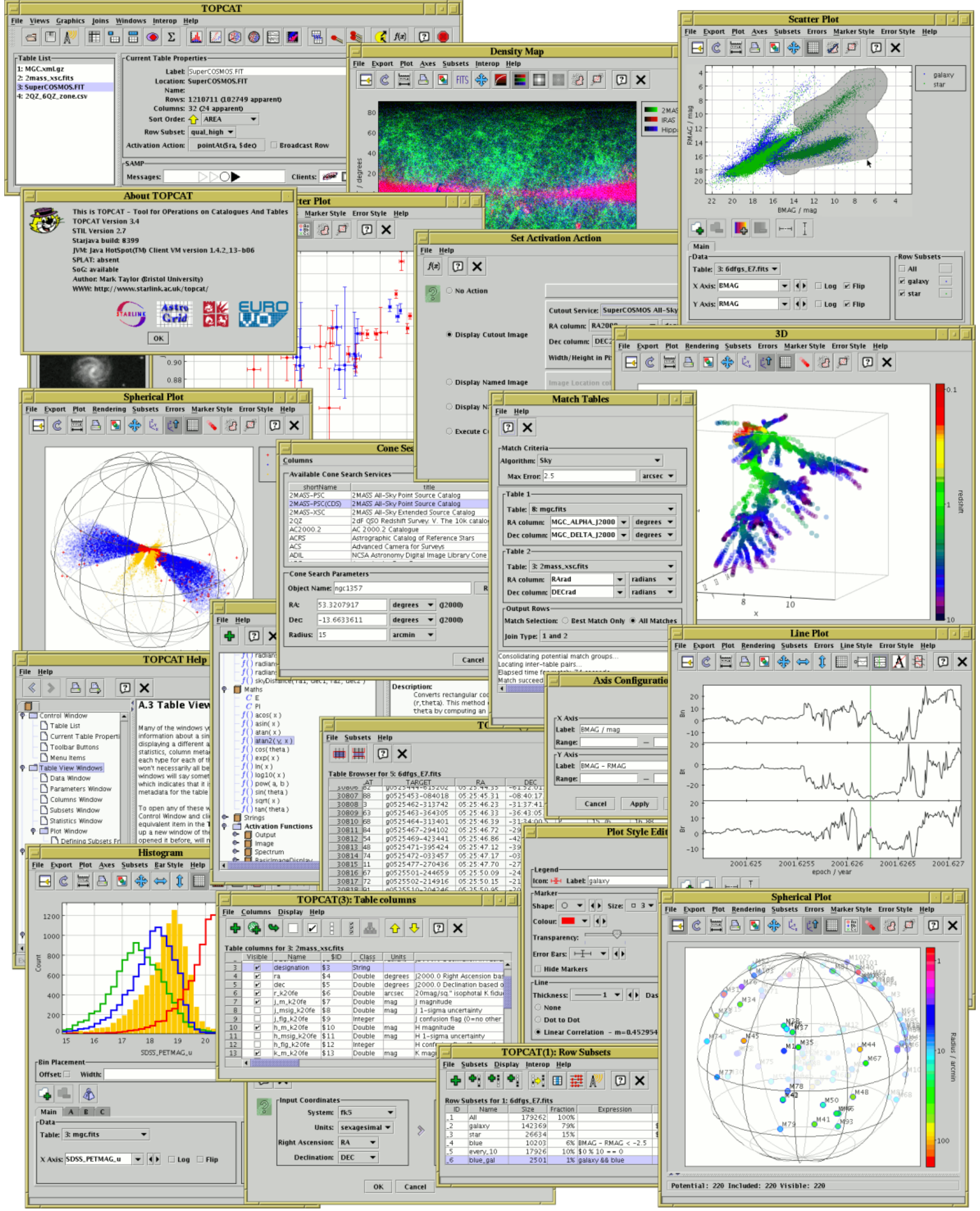
Virtualisation - building up data types…



Additional Representation Information e.g. xquery built on DRB descripion allows one to construct higher level structures

Data value obtained using Representation Information including Data Descriptions: e.g. EAST or DRB or .. Data Dictionaries: e.g. Onologies or DEDSL or …

- Building on simple virtualisation of tabular data into JTable based application
- Can combine data in different formats

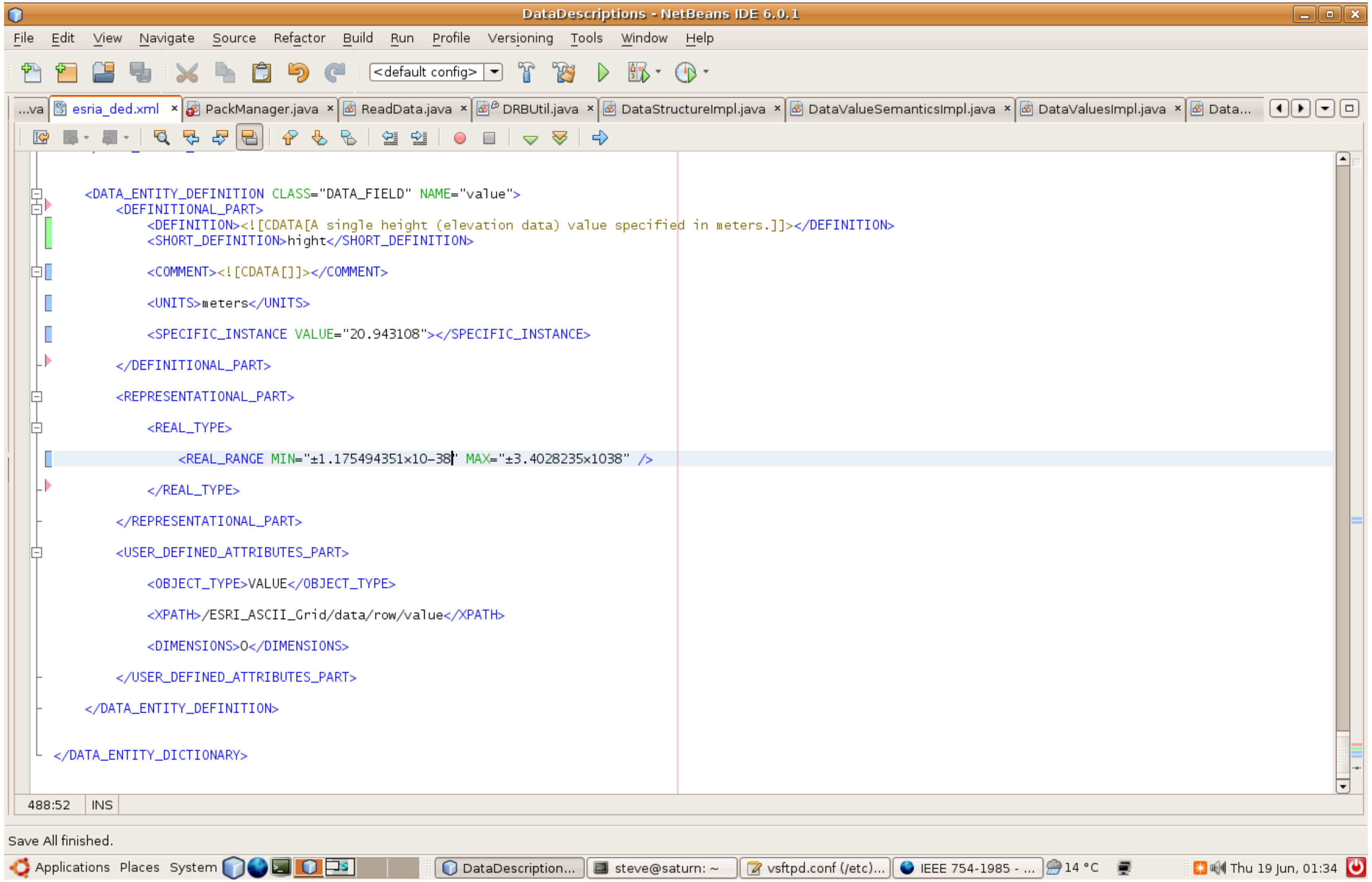


This is TOPCAT - Tool for OPerations on Catalogues And Tables
TOPCAT Version 3.4
STIL Version 2.7
Starjava build: 8399
JVM: Java HotSpot(TM) Client VM version 1.4.2_13-b06
SPLAT: absent
SoG: available
Author: Mark Taylor (Bristol University)
WWW: http://www.starlink.ac.uk/topcat/

Advantages of Formal Descriptions of Semantics

- Semantic properties such as name, description and units add meaning to the data values and data objects.
- Object oriented view of the data combined with the structure and semantics virtualises the process of going from the bits to a usable data object in software.
- Combined with formal structure it encapsulates all the knowledge needed to go from the bits to a data object in software.
- Remove the need for data specific Access Software!

DEDSL Example

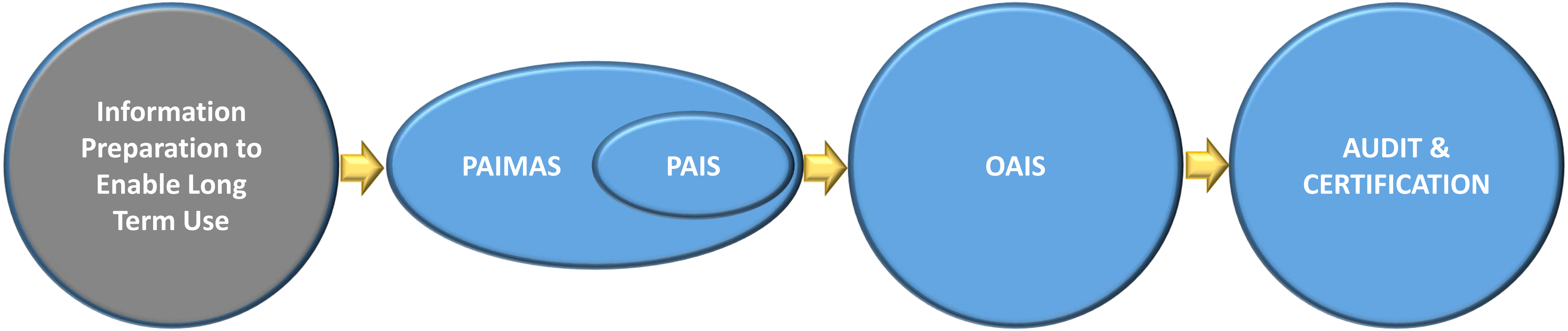


```
    <DATA_ENTITY_DEFINITION CLASS="DATA_FIELD" NAME="value">
        <DEFINITIONAL_PART>
            <DEFINITION><![CDATA[A single height (elevation data) value specified in meters.]]></DEFINITION>
            <SHORT_DEFINITION>hight</SHORT_DEFINITION>

            <COMMENT><![CDATA[]]></COMMENT>

            <UNITS>meters</UNITS>

            <SPECIFIC_INSTANCE VALUE="20.943108"></SPECIFIC_INSTANCE>

        </DEFINITIONAL_PART>

        <REPRESENTATIONAL_PART>

            <REAL_TYPE>

                <REAL_RANGE MIN="±1.175494351x10-38" MAX="±3.4028235x1038" />

            </REAL_TYPE>

        </REPRESENTATIONAL_PART>

        <USER_DEFINED_ATTRIBUTES_PART>

            <OBJECT_TYPE>VALUE</OBJECT_TYPE>

            <XPATH>/ESRI_ASCII_Grid/data/row/value</XPATH>

            <DIMENSIONS>0</DIMENSIONS>

        </USER_DEFINED_ATTRIBUTES_PART>

    </DATA_ENTITY_DEFINITION>


</DATA_ENTITY_DICTIONARY>
```

Save All finished.

New standard in development

Information Preparation to Enable Long Term Use

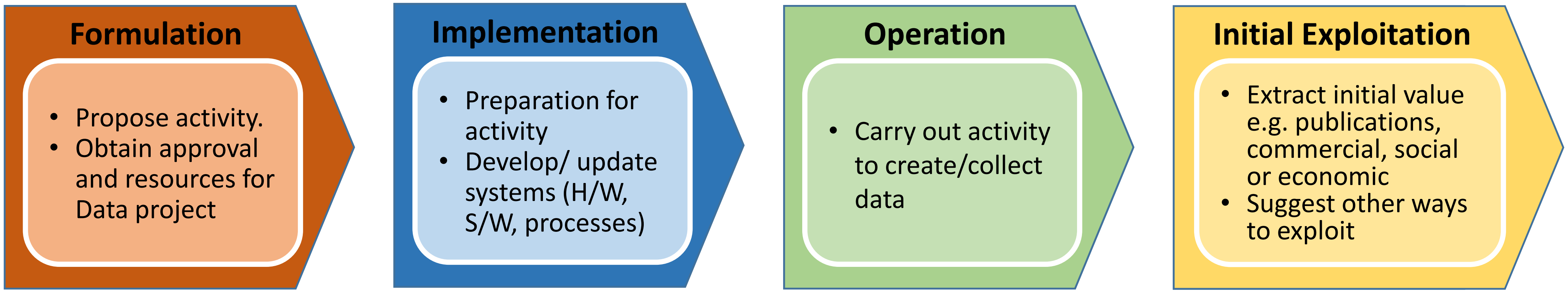


Helps to ensure that the Additional Information needed is collected/ created

Defines a mechanism to transfer the Primary Information and the Additional Information to the archive

Defines how the information should be preserved

Defines how to check that the information is being preserved



Formulation

- Propose activity.
- Obtain approval and resources for Data project

Implementation

- Preparation for activity
- Develop/ update systems (H/W, S/W, processes)

Operation

- Carry out activity to create/collect data

Initial Exploitation

- Extract initial value e.g. publications, commercial, social or economic
- Suggest other ways to exploit

Important for the development of Data Management Plans

- Working with RDA Active Data Management Plans Interest Group

Brown Dog

- Presentation from Mark Conrad

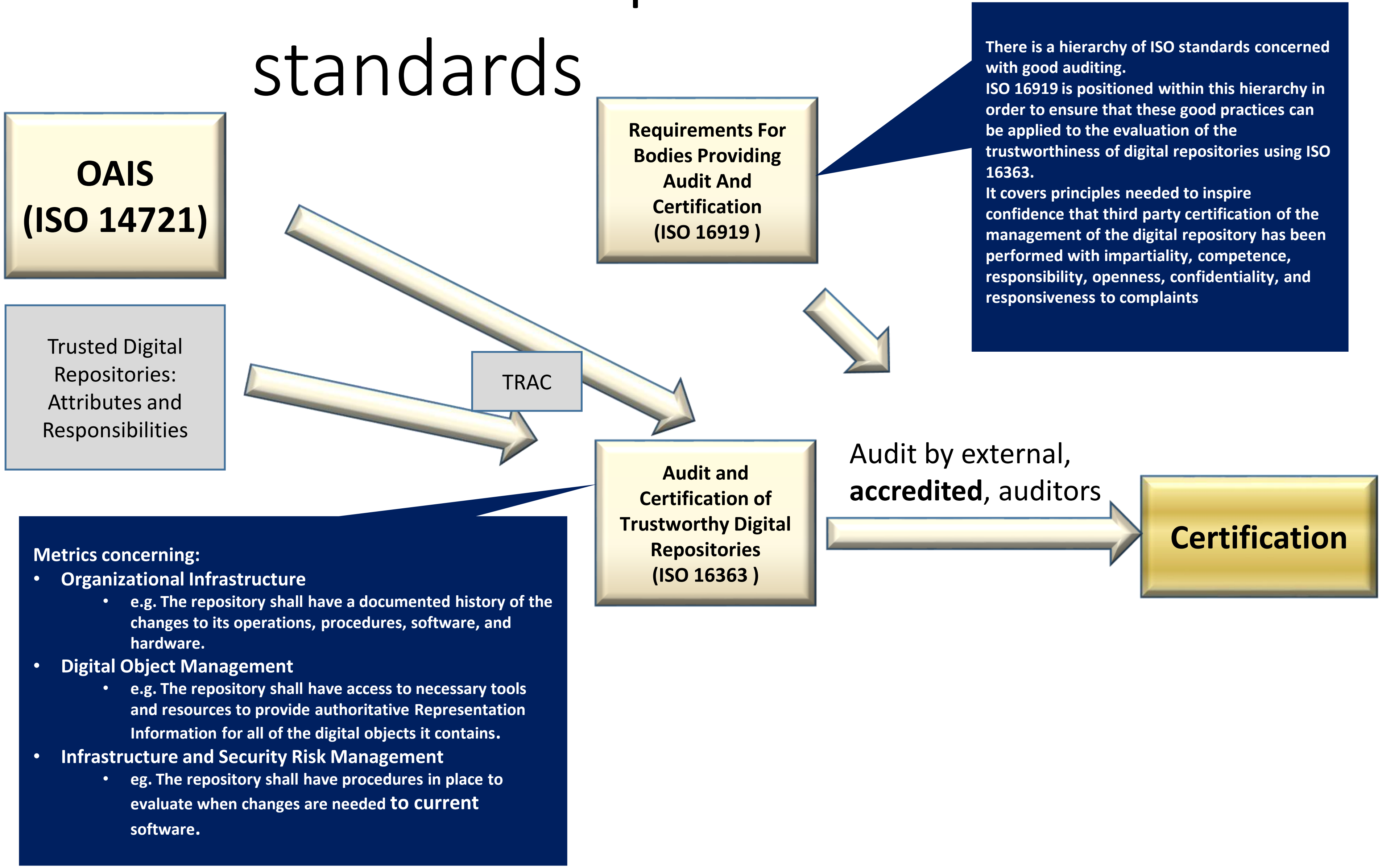
Advocating

CCSDS

- Presentations in many fora
- Working with
  - National Accreditation Bodies e.g. ANAB
  - Certification Bodies
  - various industries e.g. aerospace
  - …..
- Need help

Backup Slides

Relationship between standards



OAIS (ISO 14721)

Trusted Digital Repositories: Attributes and Responsibilities

TRAC

Requirements For Bodies Providing Audit And Certification (ISO 16919)

There is a hierarchy of ISO standards concerned with good auditing.
ISO 16919 is positioned within this hierarchy in order to ensure that these good practices can be applied to the evaluation of the trustworthiness of digital repositories using ISO 16363.
It covers principles needed to inspire confidence that third party certification of the management of the digital repository has been performed with impartiality, competence, responsibility, openness, confidentiality, and responsiveness to complaints

Audit and Certification of Trustworthy Digital Repositories (ISO 16363)

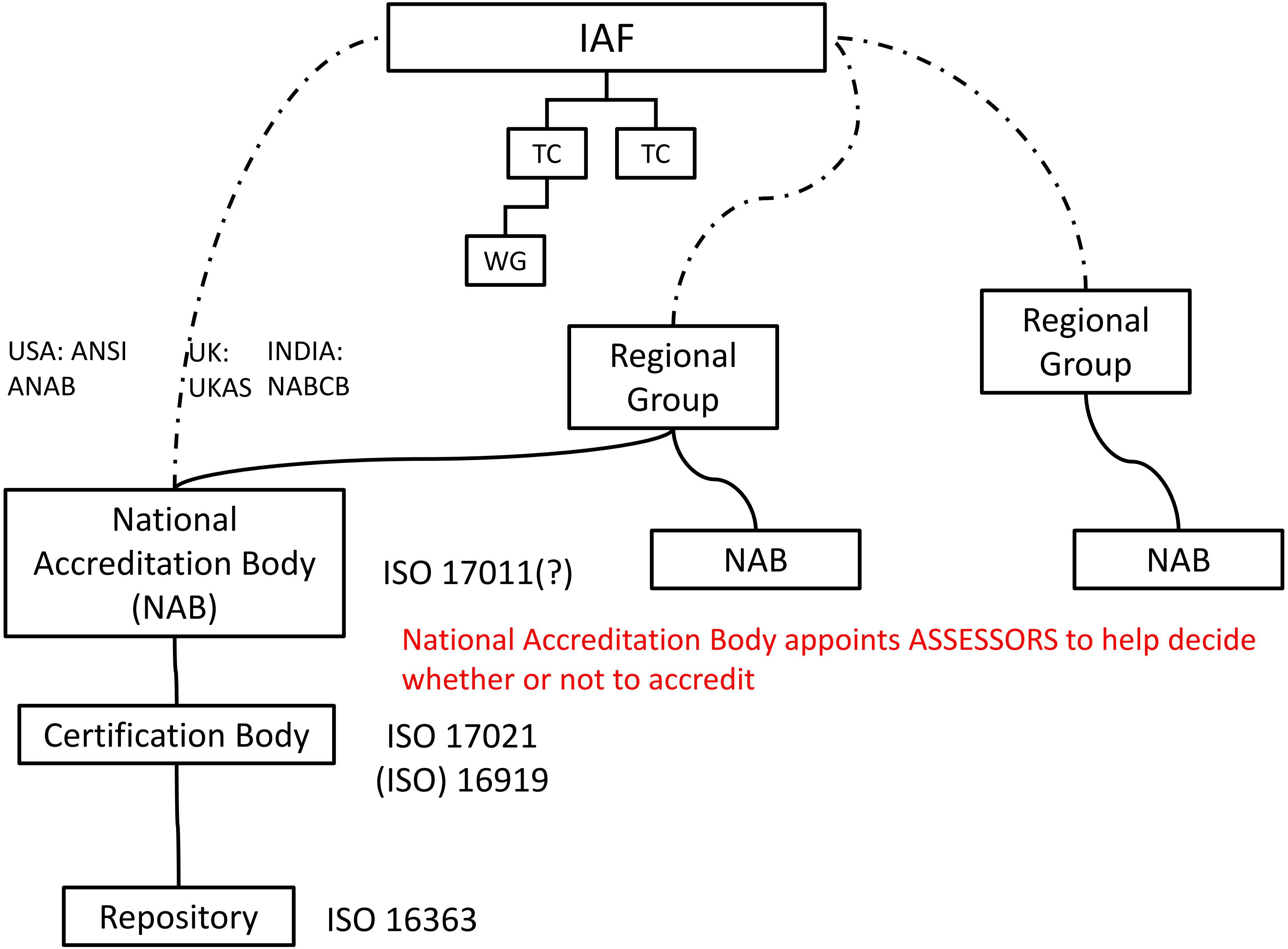
Audit by external, **accredited**, auditors

Certification

Metrics concerning:

- Organizational Infrastructure
  - e.g. The repository shall have a documented history of the changes to its operations, procedures, software, and hardware.
- Digital Object Management
  - e.g. The repository shall have access to necessary tools and resources to provide authoritative Representation Information for all of the digital objects it contains.
- Infrastructure and Security Risk Management
  - eg. The repository shall have procedures in place to evaluate when changes are needed to current software.

See http://wiki.digitalrepositoryauditandcertification.org and
http://www.alliancepermanentaccess.org/membership/member-resources/audit-and-certification
Standards will be available free from http://www.ccsds.org



National Accreditation Body appoints ASSESSORS to help decide whether or not to accredit