Data Management Server Presentation

Peter Miu

20th February, 2008

EUMETSAT

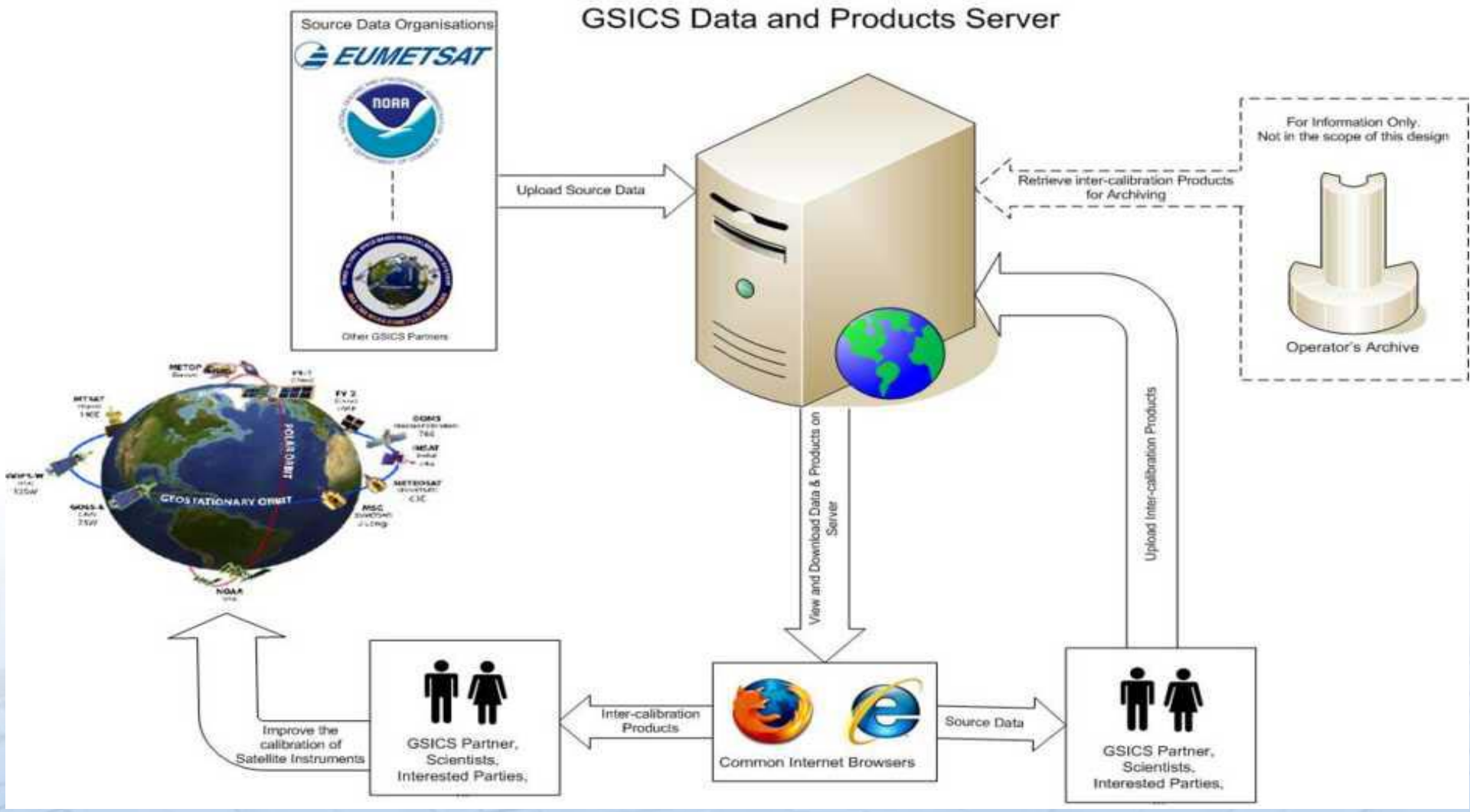
Purpose of the Presentation

- Propose the requirements for the data management server to support GSICS.
- Use this as the basis for discussion for the implementation of a data management server.
- Finalise on a set of agreements between the GSICS partners on this implementation.

Proposed User Requirements

- The data management server shall allow GSICS partners to upload source data and products into a ‘rolling archive’ for use by other GSICS partners.
- The source data and products shall be presented in a webpage and is downloadable via a web-browser.
- The operator of the server can optionally archive the products received.
  - This is recommended for data preservation and it will benefit a wider user community.
  - Whether archiving requirements falls under the implementation of the data manager server is up for debate as this is not specified in the primary goals of GSICS (see http://www.orbit.nesdis.noaa.gov/smcd/spb/calibration/icvs/GSICS/index.html)

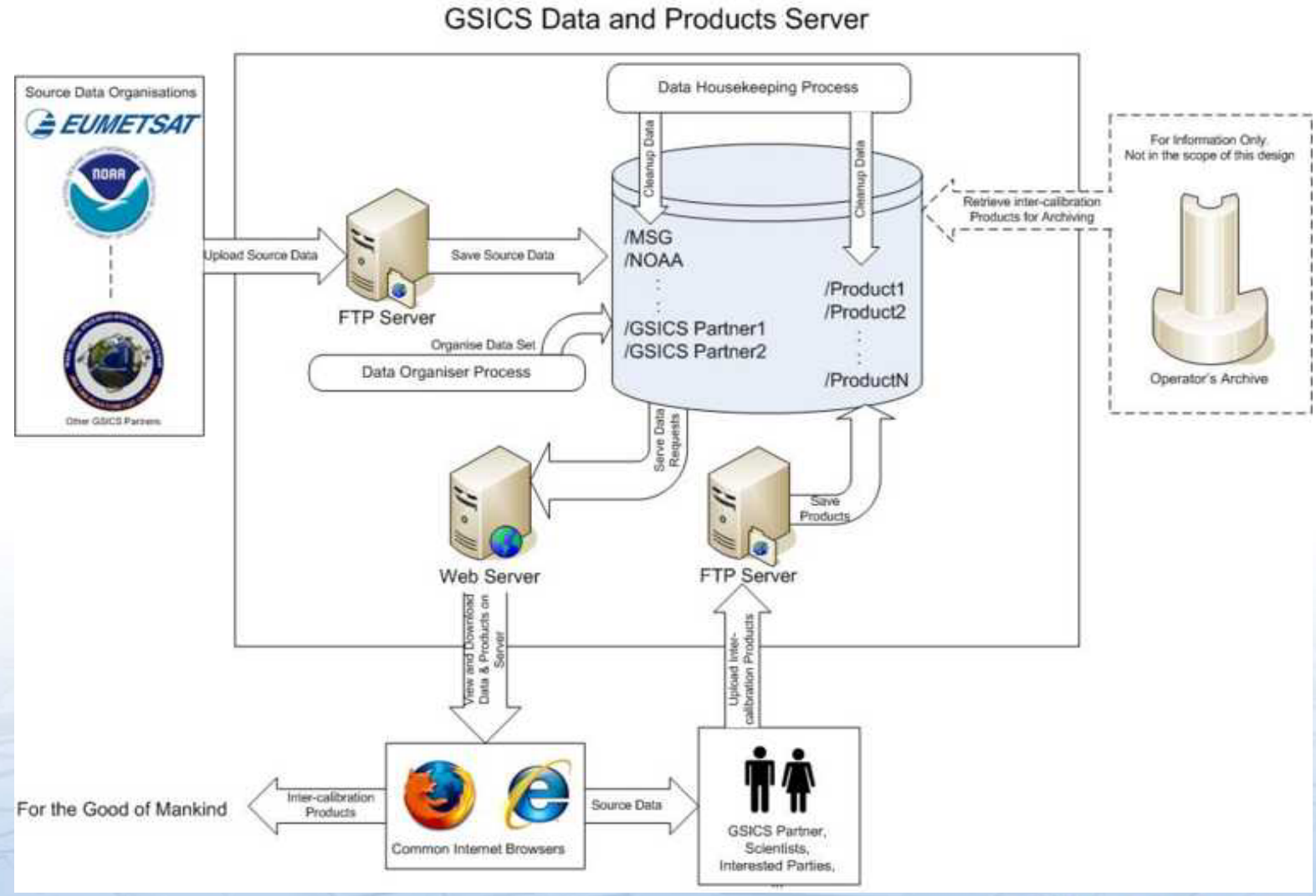
Generic view of the Data Management Server



Basic Specification for the Data Management Server

- High end server with enough memory, and disk space to accommodate the source data and products from the GSICS partners (Data flows and formats have been agreed).
- FTP server installed.
- Web server installed.
- Processes for:
  - Organising the Data and Products received in a form accessible via the internet.
  - Housekeeping this Data and Products to maintain a rolling archive.

Revised View of the proposed Data Management Server



Proposed Data Management Server Interfaces

- For uploading source data from GSICS partners:
  - Standard FTP user account; one account per GSICS partner. Predetermined source data are processed. Unknown data formats received are deleted.
- For uploading products :
  - Standard FTP user account; anonymous FTP user, upload to predefined directories. Product authors are required to follow agreed standard for organising the product's meta data. Unknown data formats are deleted.
  - Or development of a sophisticated web service client that connects to a web service running on the data management server. This client can structure products by adding mandatory meta data information before uploading the product to the data management server.

Data Organiser Process

- The process shall:
  - Identify the source data or products received.
    - Using agreed filenames.
    - Using agreed file contents contain meta data.
  - Organise this data in a manner that can be accessible and downloadable by the web server.
  - Refresh the web pages generated once house keeping has been perform on the ‘rolling archive’.
- The Unidata THREDDS data server is a supported COTS that can be easily configured as a data organiser process.

Housekeeping Process

- The process shall:
  - Periodically scan the source data and products' directories for expired content. Contents have expired when:
    - Their shelf life exceeds pre-defined number of days from the content's reception date.
      - Each directory can have a different shelf life defined.
    - Overall disk available percentage resource of the server has reached a threshold level indicating no further content can be accommodated.
      - A global disk available percentage threshold is defined. Individual disk percentage threshold can also be defined for each directory.
- Development of a housekeeping process is relatively simple. If the server is a Unix platform, the crontab daemon can be used to perform housekeeping via the corresponding shell commands (find and rm).

nter-Operability Issue

- Inter-Operability is defined to be :
  - The ability of a system to work with, or use the parts or equipment of another system.
- For the GSICS Data Management Server, the following components shall be inter-operable.
  - Source data formats.
  - Product formats.

nter-Operability : Source Data and Products Identification

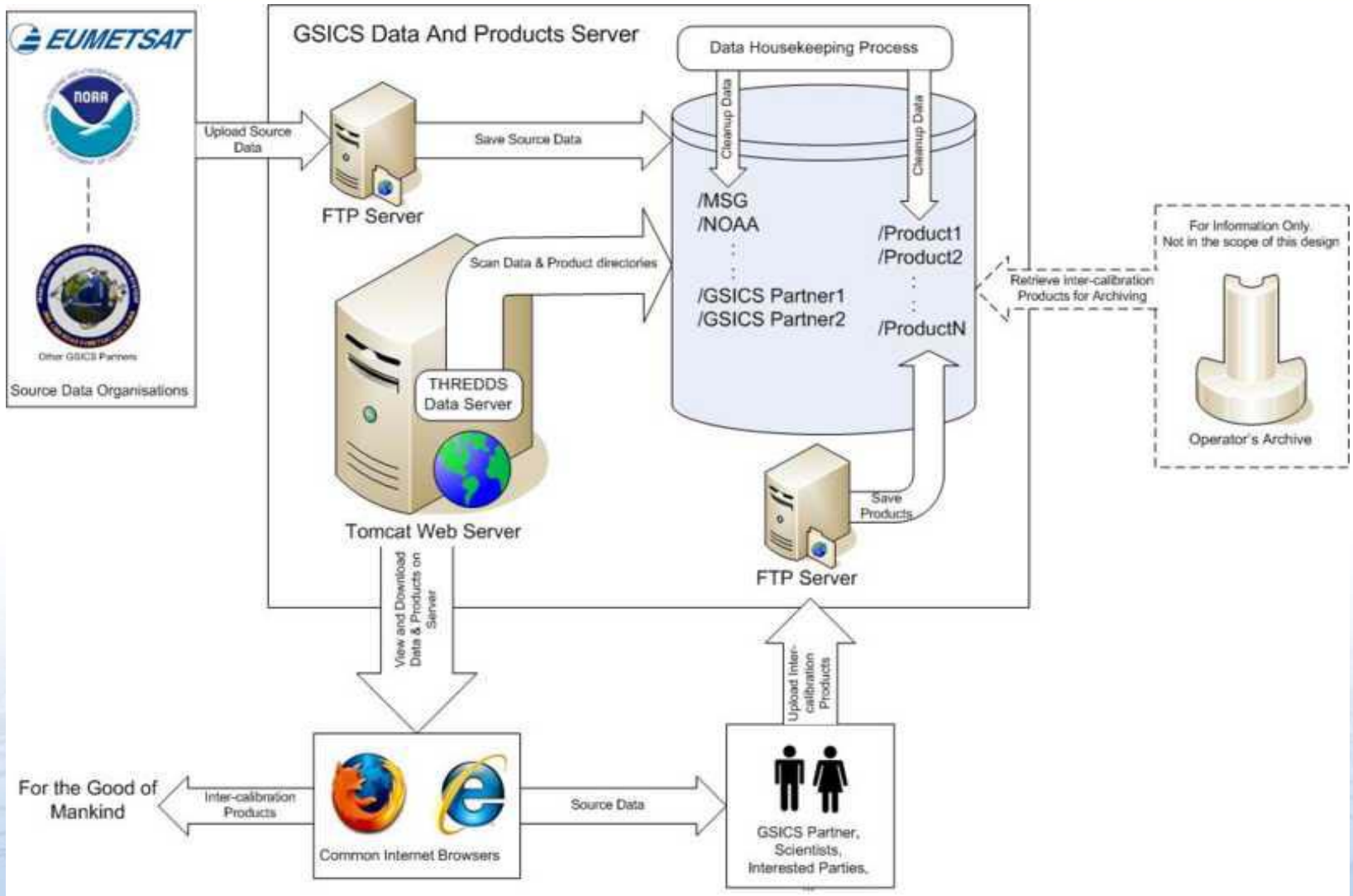
- By an agreed Filename Format.
  - Advantages: simple, human readable.
  - Disadvantages: prone to error, no guarantee of file content.
- By agreed File Content.
  - Advantages: relatively simple, less prone to error.
  - Disadvantages: each new format requires an implementation to identify it.
- By agreed Meta Data
  - Advantages: no or minimal implementation updates are needed to support new source data and products.
  - Disadvantages: no guarantee of actual file content, meta data file may be lost during transfers.
- Using a Data Format containing Meta Data in its Structure
  - Advantages: simple once agreed mandatory meta data fields are defined.
  - Disadvantages: resultant files can be big.

nter-Operability : Source Data and Products Identification Recommendations

- Source data and scientific products should be in a data format containing agreed meta data values.
  - A possible candidate for this data format is NetCDF containing Climate Forecast Metadata Convention.

- Other types of products; algorithms, source code, executables, documentation etc. should follow the filename format convention for identification.

Revised View of the proposed Data Management Server



nter-Operability : Archiving

- Should archiving the products received on the data management server be an inter-operability issue to be agreed on under the scope of the data management server implementation ???

- Should this be left to the operator of the server to decide whether and how to archive this data ??

Advantages of the Proposed Implementation

- The advantages of this design are:
  - Utilises free COTS supported by a funded organisation.
  - Relatively simple to implement as a pilot version of a server in the federation of GSICS FTP servers.
  - Inter-operability standards for identifying data sets with a view to archiving, querying and retrieving products are encapsulated between the server and the Operator's Archive. The complexity of various generic standards are removed from the creators of the inter-calibration products. The inter-operability issue should be part of the Operator's Archive rather than a requirement of the GSICS Data and Products Server's 'rolling archive'.

Specification of Deliverables for the Data Management Server

- Server design; hardware, software, dataflow specification i.e.
  - What data.
  - Frequency the data is expected.
  - Shelf life of the data; how long the data shall exist on the server.
- Source data format specifications including filename formats.
- Products specifications including filename formats.
  - Proposed format of the product, software, documents up-loadable back to the server.
  - Shelf life of these products.
- Archiving recommendations (?).

Tentative Implementation Schedule

- Implementing a EUMETSAT's version of the proposed server can be achieved by the 4th quarter of 2008.
- Once the implementation has been completed. The operational aspects can be analysed and recorded. Evolution of the server can then be discussed and progressed.
- Report of the findings will be presented in future GSCIS working group meetings.

END OF PRESENTATION