Outcomes of the CDS Technical Infrastructure Workshop

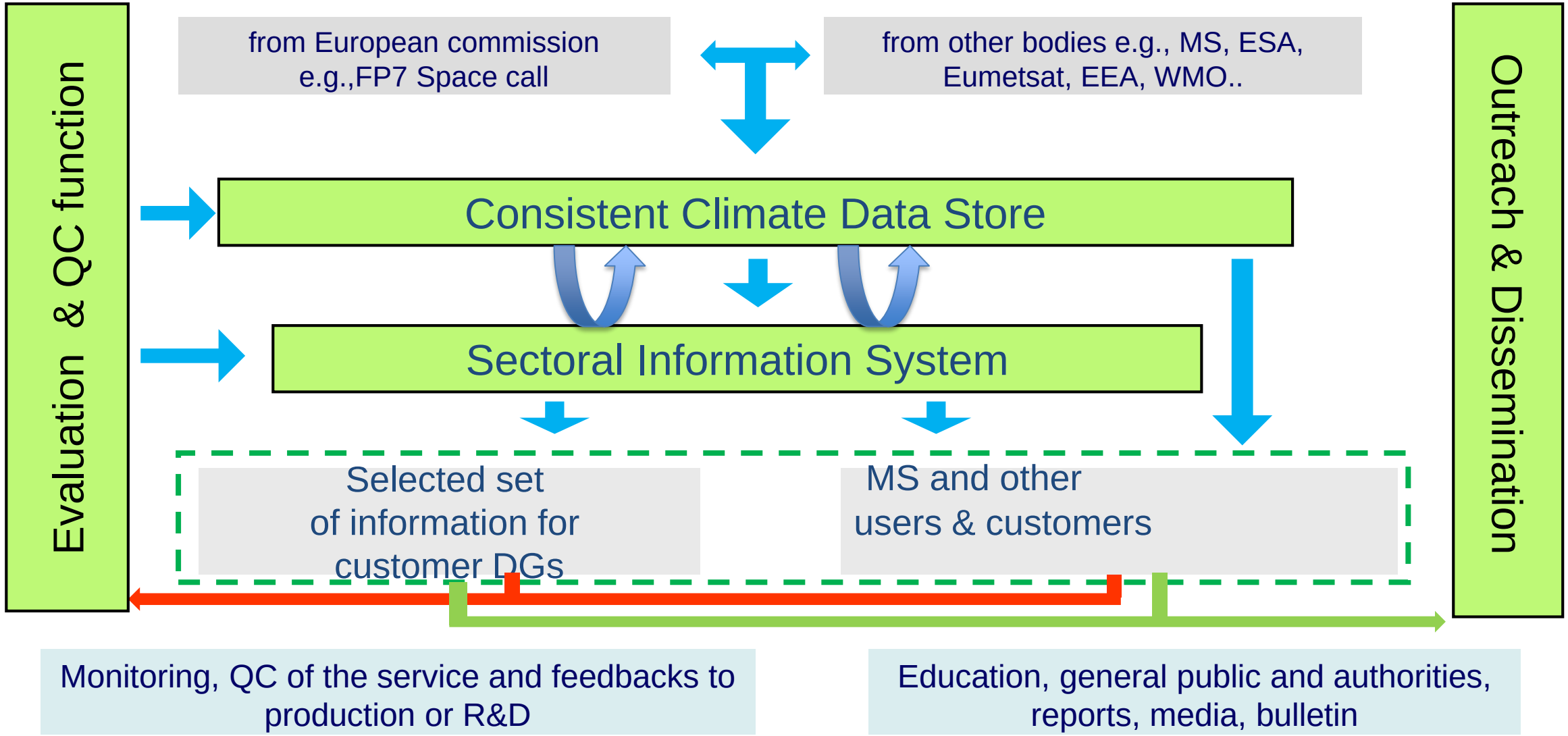
Baudouin Raoult

Baudouin.raoult@ecmwf.int

Funded by the European Union

Implemented by ECMWF

C3S architecture



C3S Data Store

- *… the service draws upon the outcome of the FP7 Copernicus* ***precursor projects*** *…*
- *(products)… will have to be accessible in an* ***operational*** *way*
- *…technical development, maintenance and* ***governance*** *efforts will be required from the* ***data providers*** *to ensure fully* ***compliance*** *with the C3S requirements*

C3S Data Store

- *The **EQC** will … monitor … using standard **key performance indicators***
  - *… technical **quality of service** as measured by timeliness, number of interruptions, response time for troubleshooting…*
  - *…**quality of products** through statistical comparison with observed quantities;*
  - *…**quality of information** made publicly available …*
  - *…**uptake** of services and products **by users**: …unique visitors on the web portal, downloads, data volumes…*

C3S Data Store

- *…access to the products for **authenticated users** ….*
  - *… single logon across the Copernicus programme (mid-term)*
- *…identification of **backup solutions** regarding the provision of information populating the CDS and the SIS*
- *… the provision of a technical user **support** and **help desk** facility…*

C3S Data Store

- *Timely acquisition of state-of-the-art climate information from various data providers, and the development and maintenance of the **C3S catalogue** content*

- *The information delivered to the end-user is **fully traceable**, **quality controlled** and **disseminated** within the most appropriate time*

- *To ensure uptake of climate information by downstream users, **climate toolboxes** will be developed and maintained*

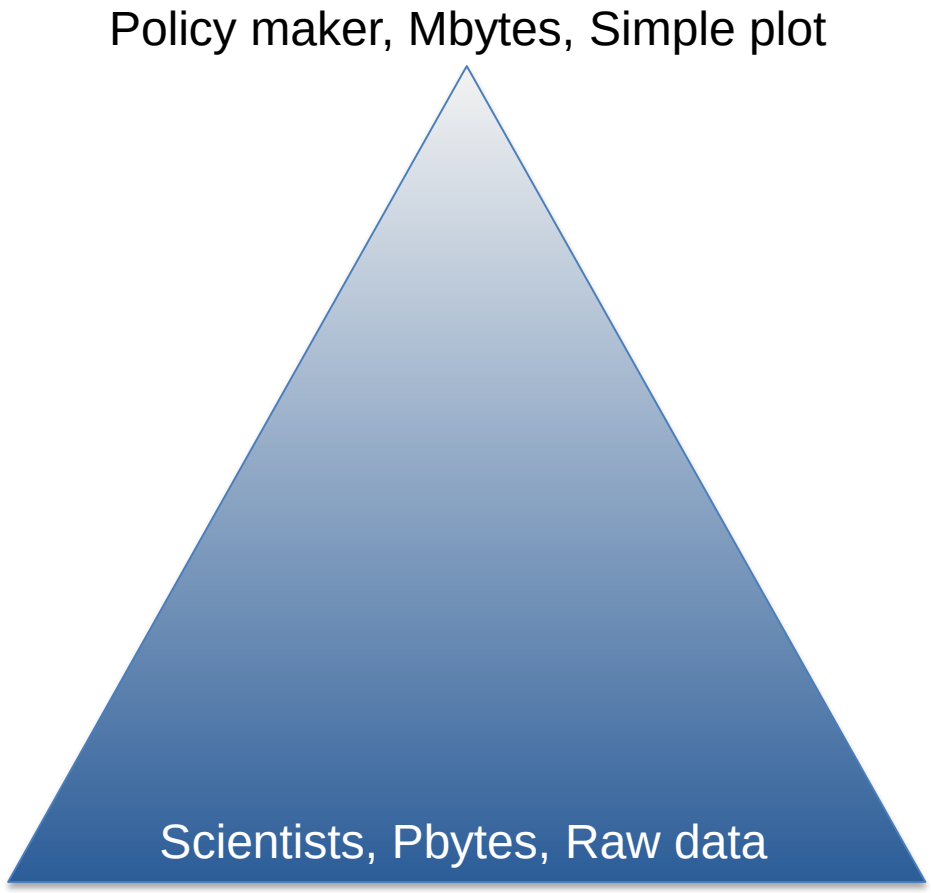
Requirements for the Climate Data Store

- Be **distributed**
- **Reuse** existing systems when possible

- ... But **should not** be a mere collection of heterogeneous systems:
  - The user should have a **consistent** view of all data and services available through the CDS

Main challenges

- Diversity of **users**
  - Scientist to policy makers
- Diversity of **volumes**
  - PB to KB
- Diversity of **products**
  - Raw to elaborated



Data (PB) → **Information** (TB) → **Knowledge** (GB) → **Wisdom** (MB)

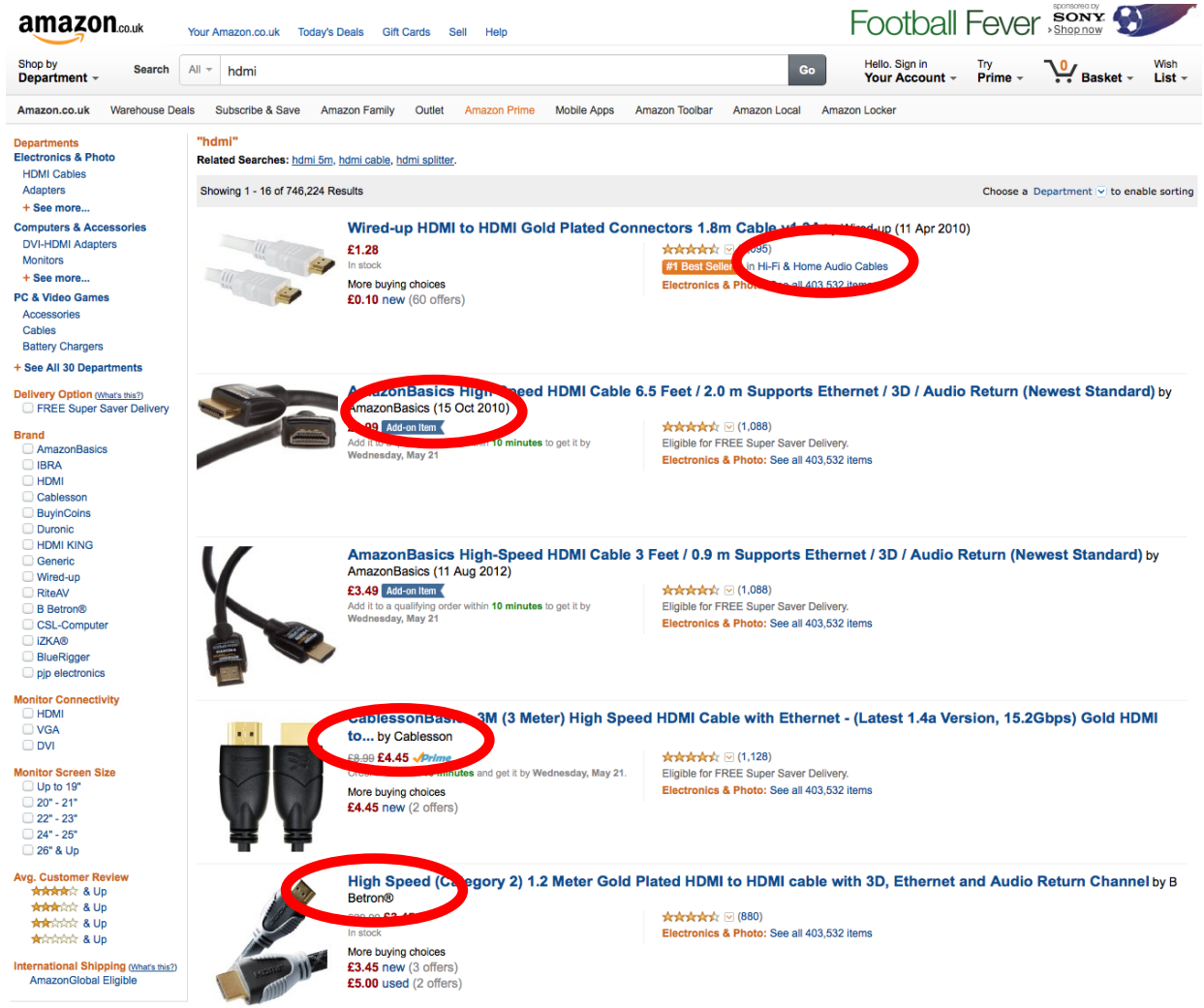
opernicus

ECMWF

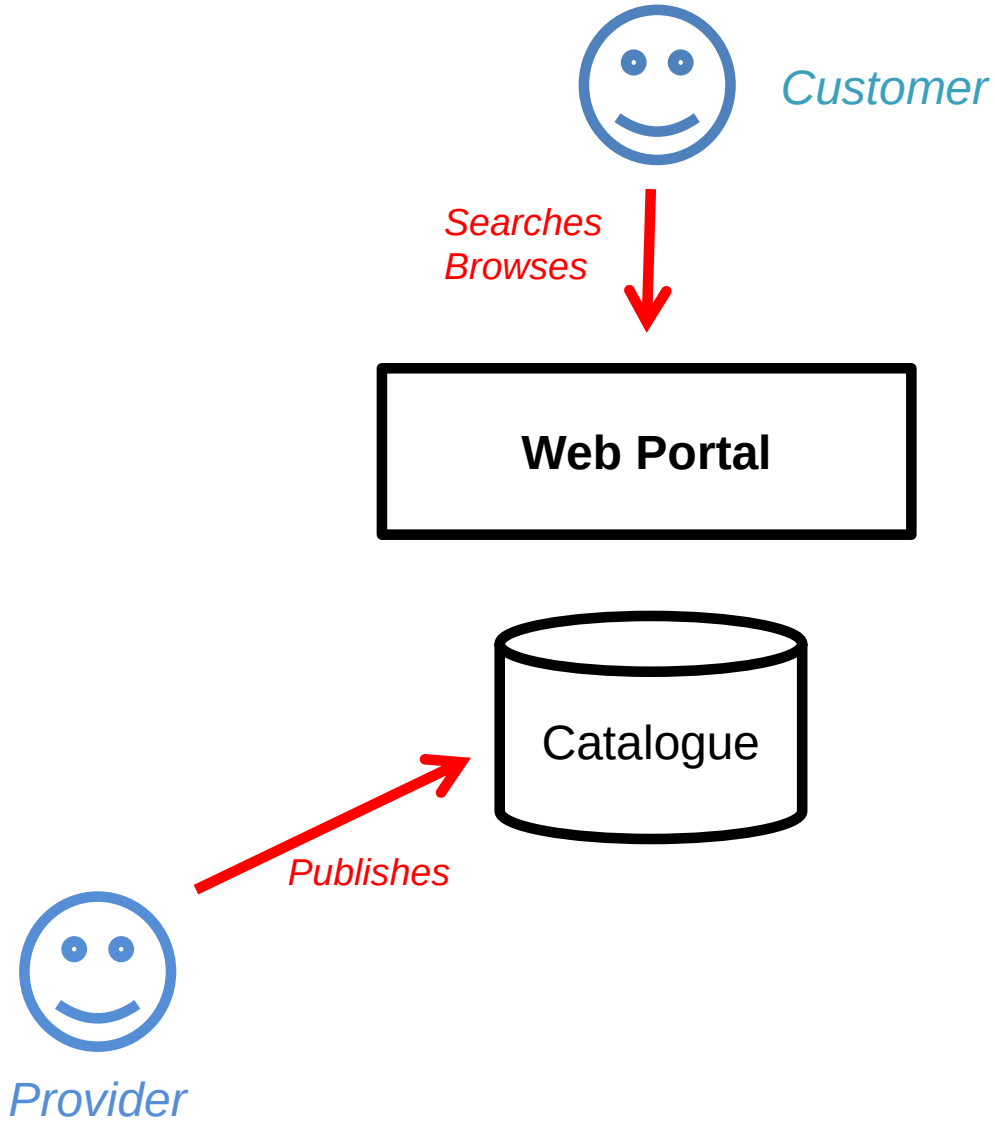
What is a PiB? *(Assuming reading from/writing to disk at 100 MiB/s)*

	Bytes	Seconds	Days	Months
MiB	1,048,576	0.01		
GiB	1,073,741,824	~10		
TiB	1,099,511,627,776	10,485	0.12	
PiB	1,125,899,906,842,624	10,737,418	124	**> 4**

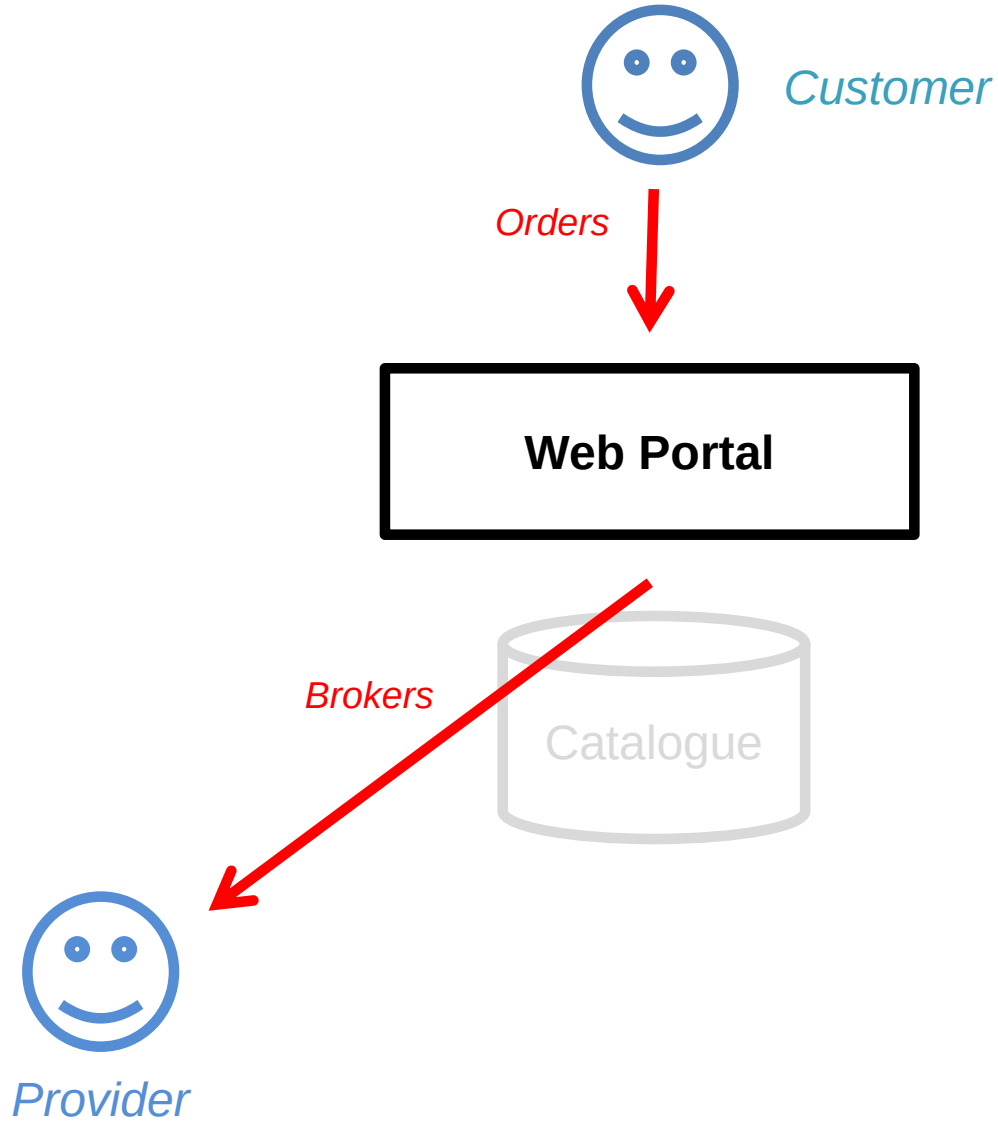
Example: Amazon marketplace



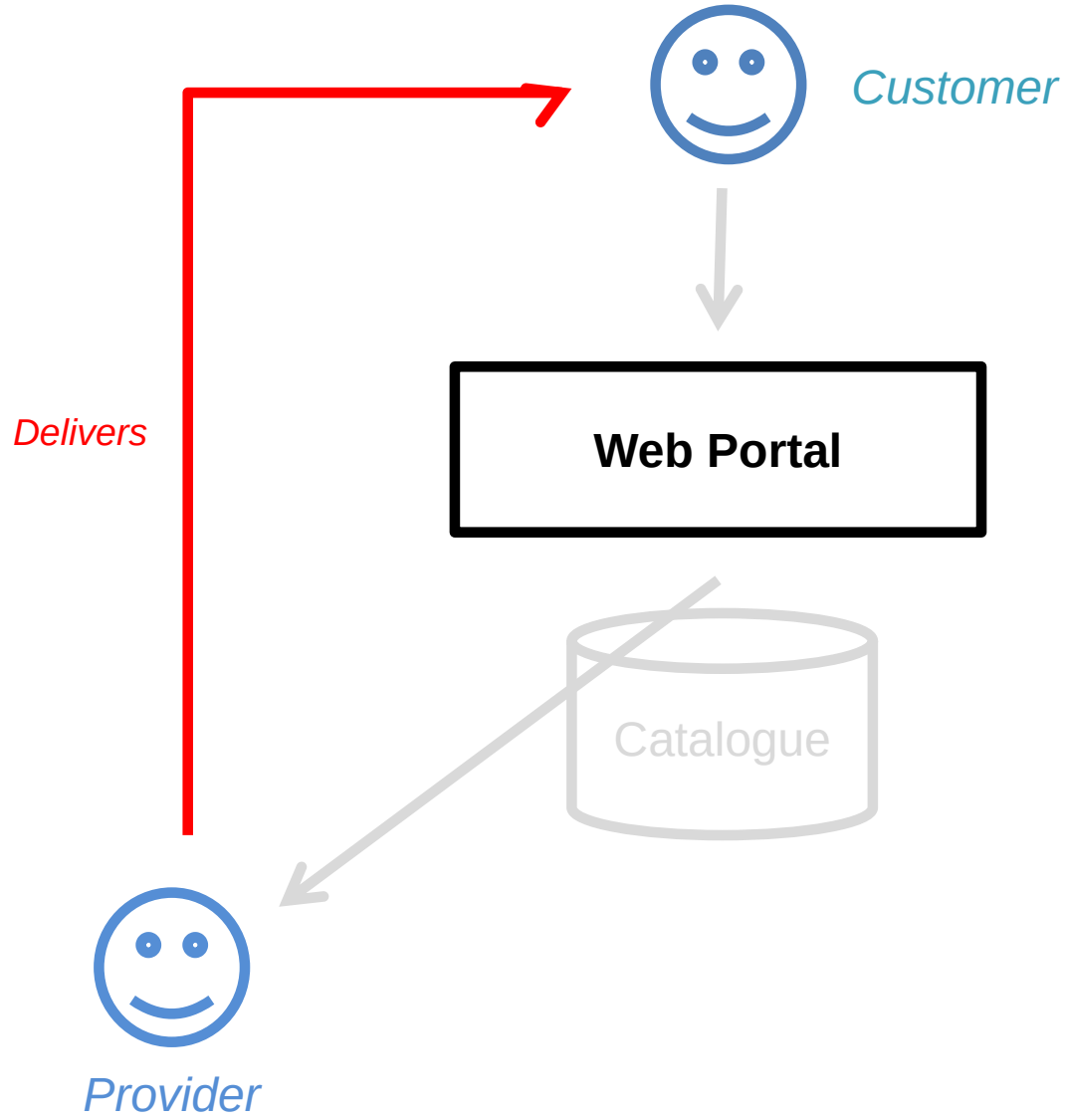
A “Marketplace”



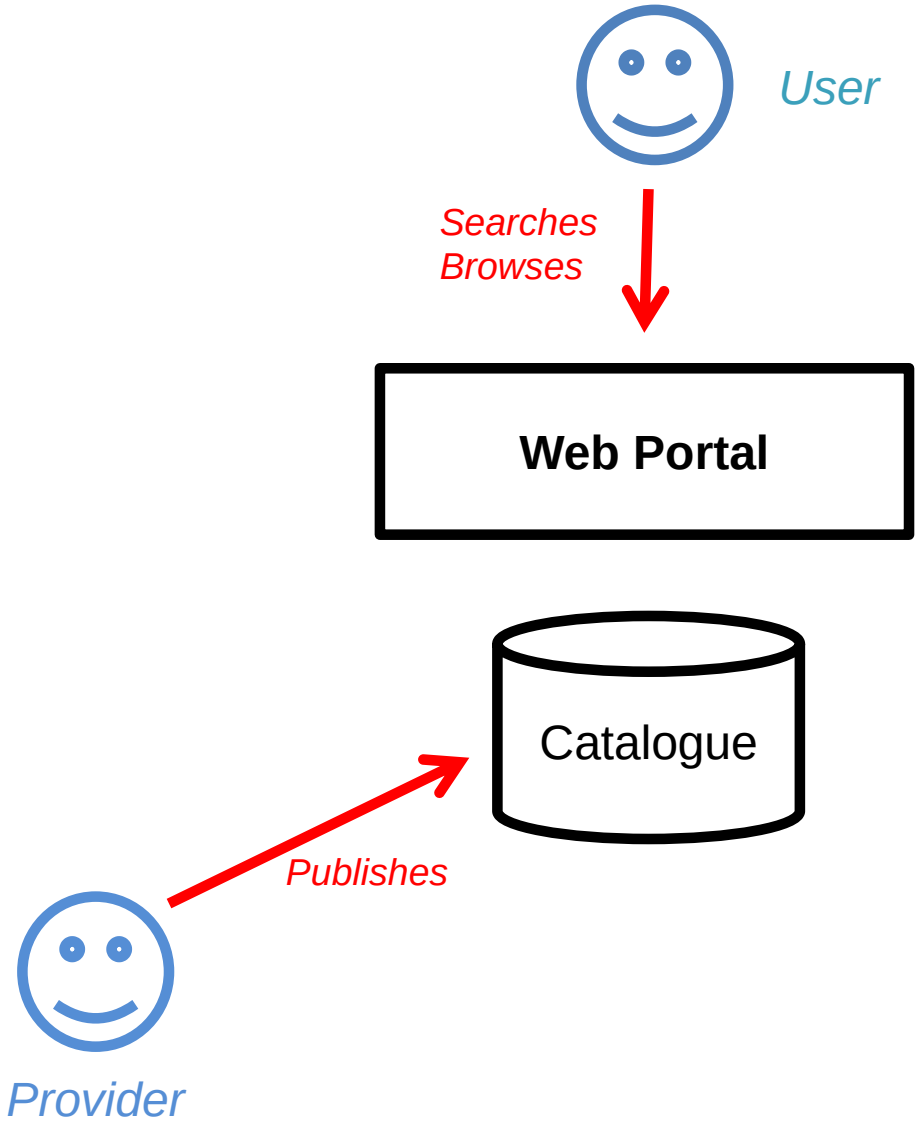
A “Marketplace”



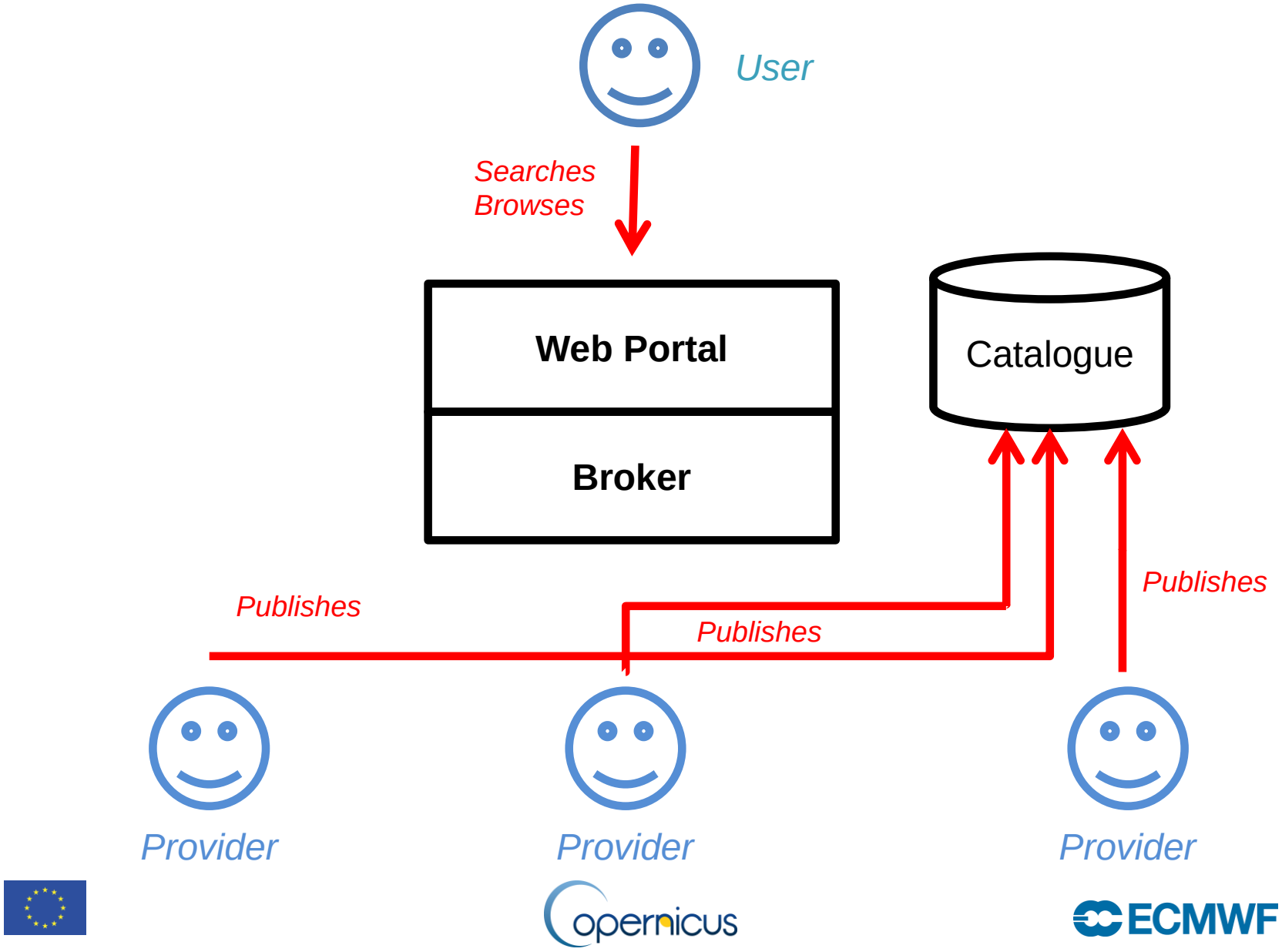
A “Marketplace”



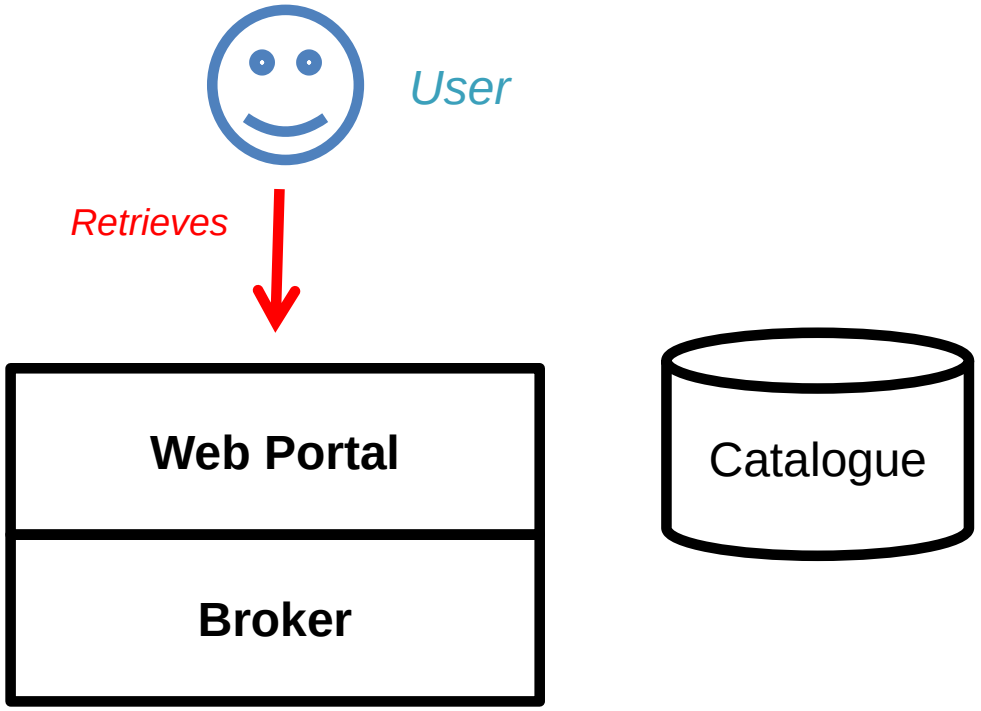
Climate Data Store: Architecture



Climate Data Store: Architecture



Climate Data Store: Architecture

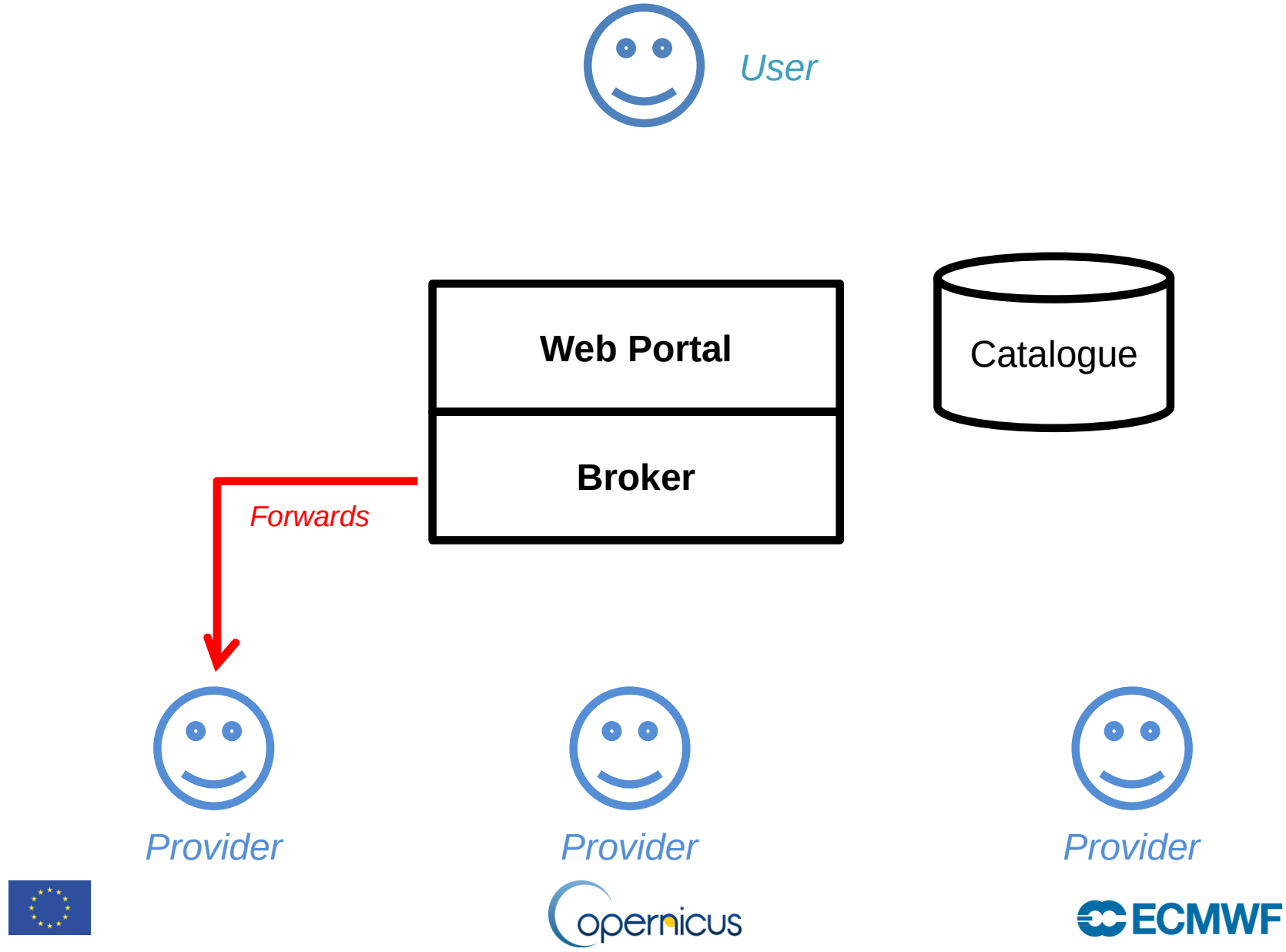


Provider

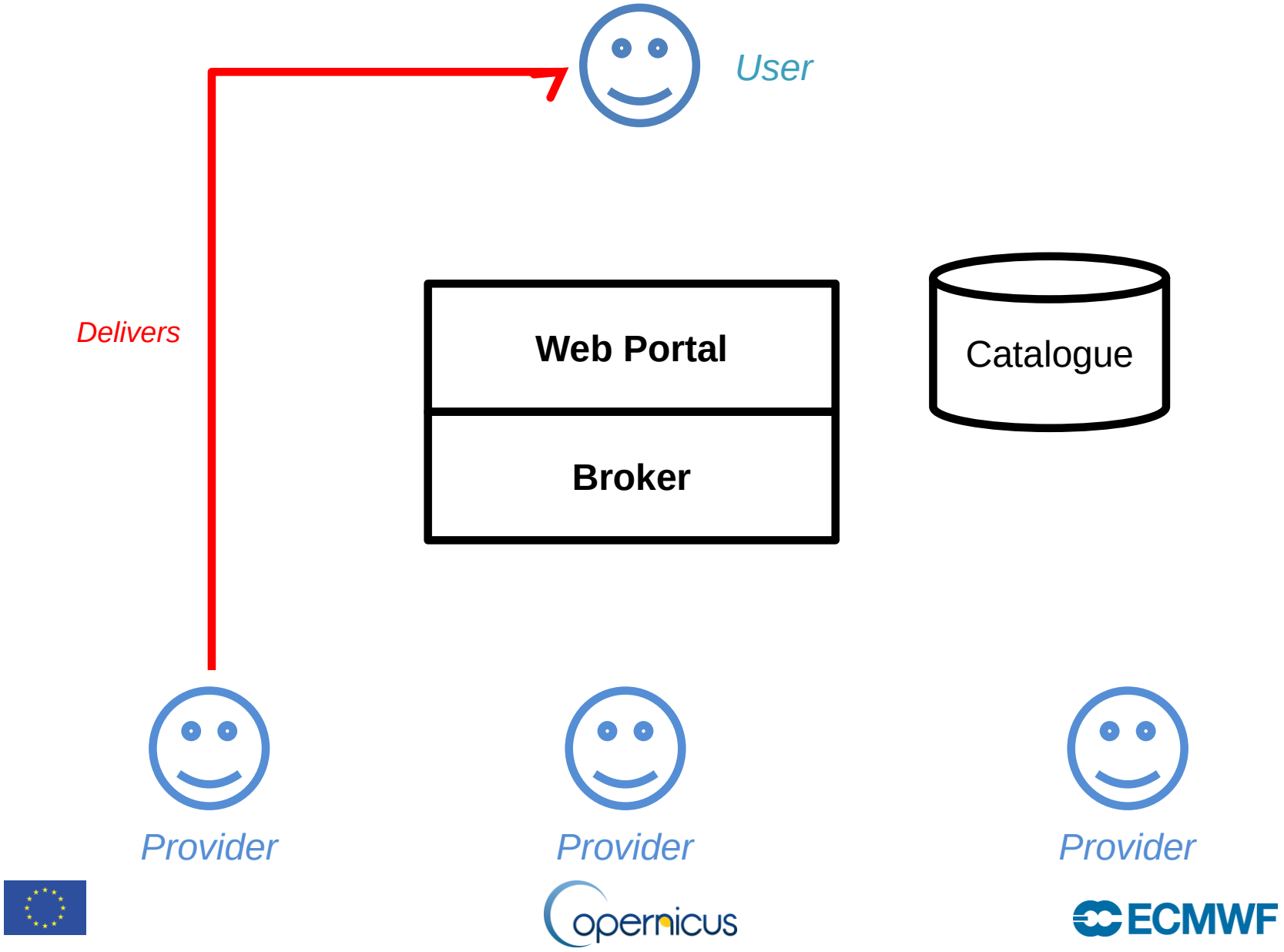
Provider

Provider

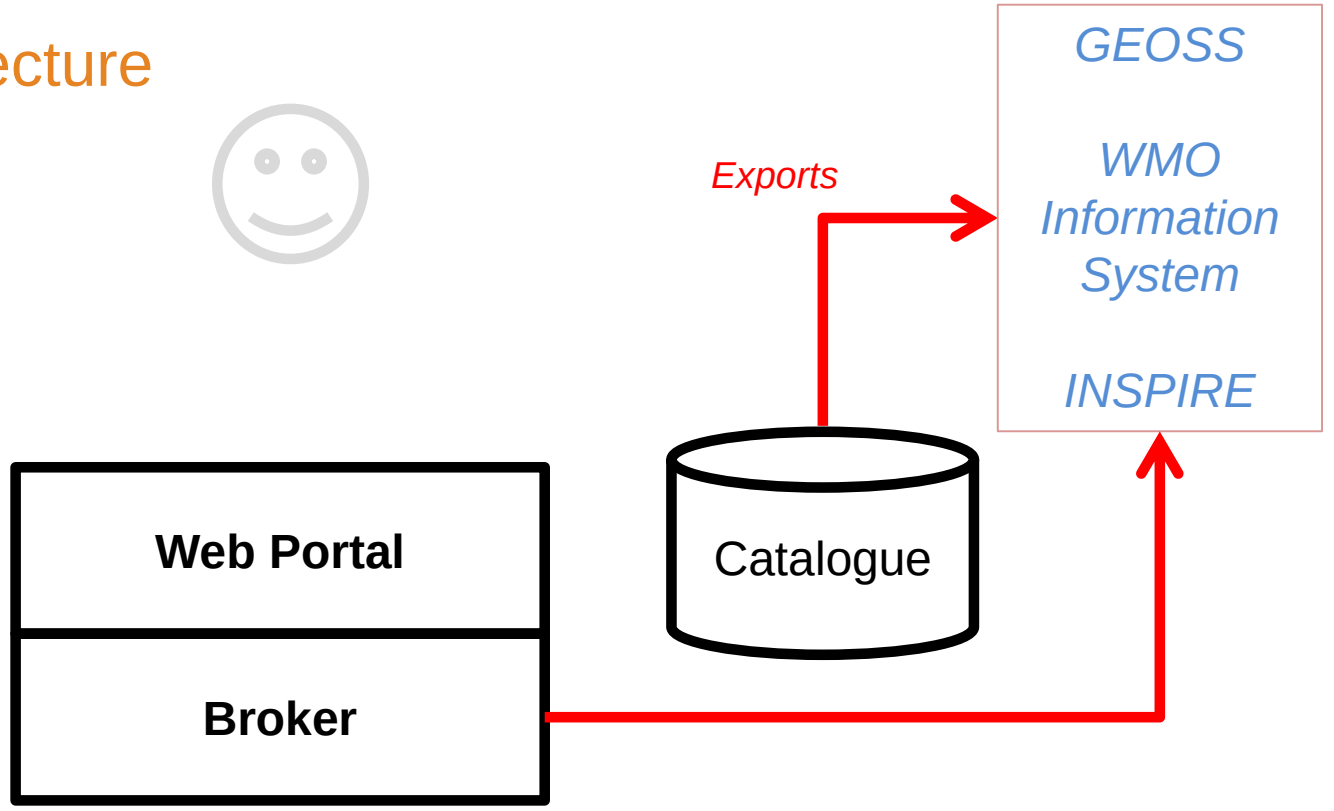
Climate Data Store: Architecture



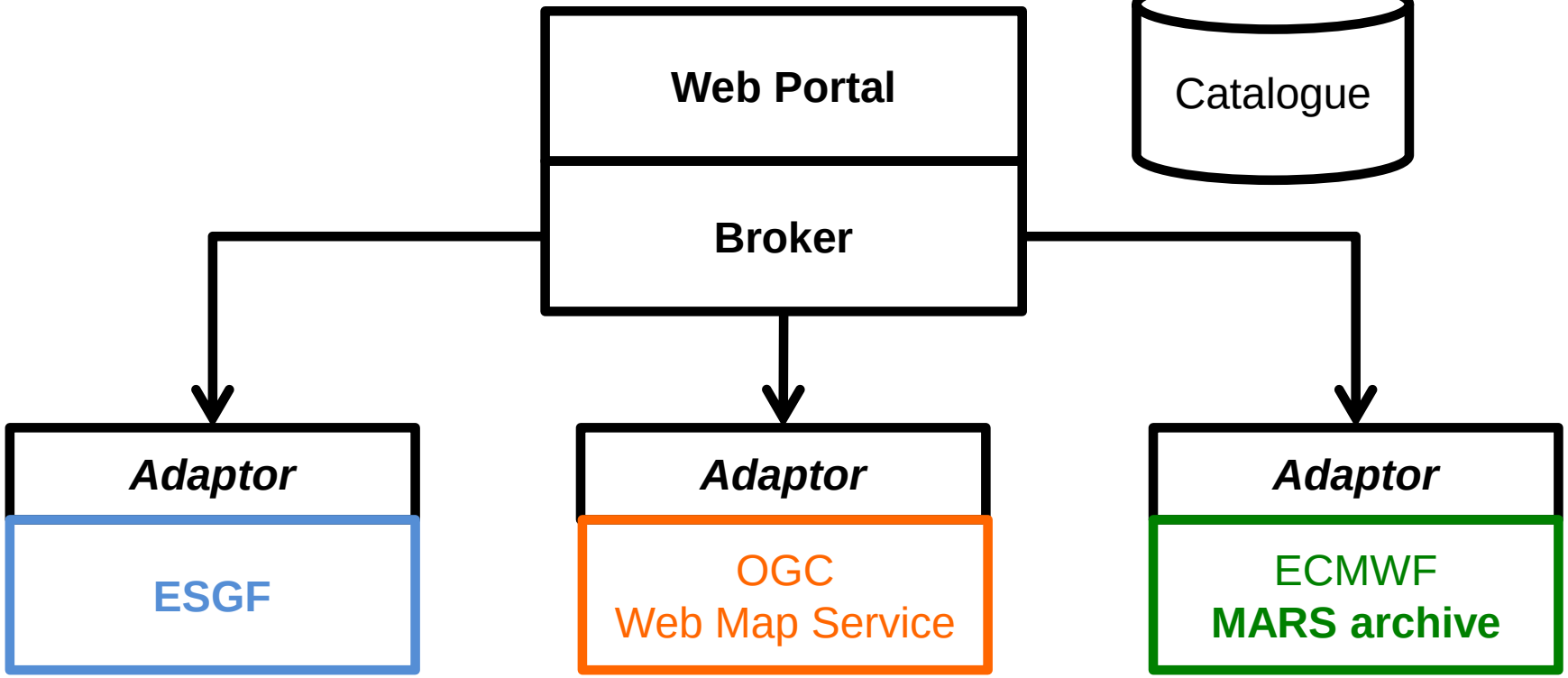
Climate Data Store: Architecture



Climate Data Store: Architecture



Climate Data Store: Architecture



Climate Data Store: Architecture

- ECMWF could host new services when **no infrastructure** exists

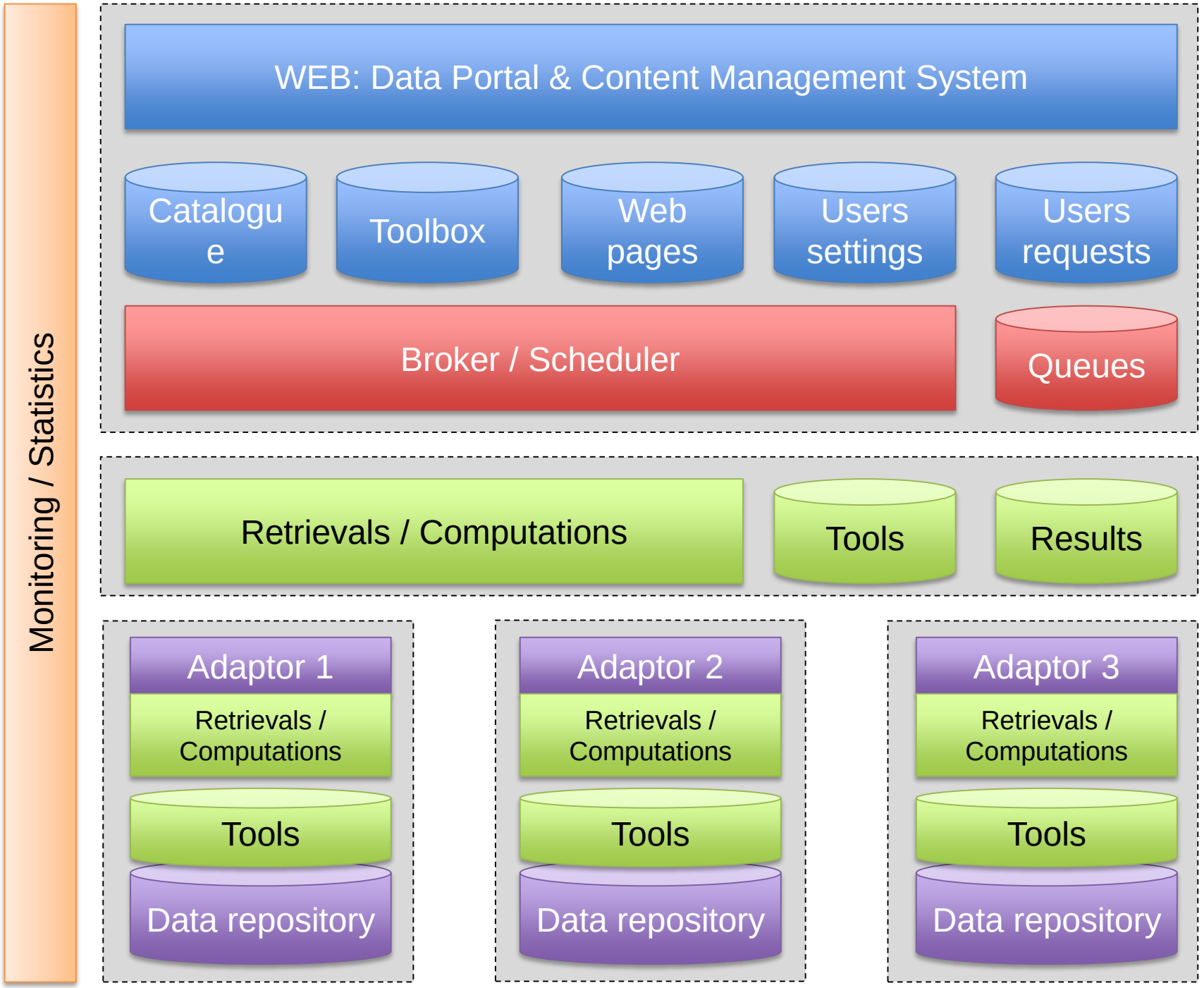
- Adaptors are **not limited** to data provision
  - They will contribute to the “C3S toolbox”
  - As for data, services are invoked by the broker

Operational?

- Monitoring
- Reporting
  - Capacity planning
  - Usage statistics

- Service level agreement
- On-call and support
- Help desk

- High-availability
- Backup


Monitoring / Statistics
WEB: Data Portal & Content Management System
Catalogue
Toolbox
Web pages
Users settings
Users requests
Broker / Scheduler
Queues
Retrievals / Computations
Tools
Results
Adaptor 1
Retrievals / Computations
Tools
Data repository
Adaptor 2
Retrievals / Computations
Tools
Data repository
Adaptor 3
Retrievals / Computations
Tools
Data repository

ECMWF CDS Workshop, 3-6 March 2015

- 70 participants: EU institutions, NMSs, research institutes and commercial companies
- 19 countries
- 40 presentations (4 via videoconferencing)

About the CDS workshop

- Aim: discuss the **development** of the C3S software **infrastructure**

- Three **themes**:
  - User expectations
  - Existing Climate Service Providers
  - Industry Perspectives

- Working **groups**
  - Catalogue and Portal
  - The 'Toolbox'
  - Content, Standards and Interoperability.

Findings and recommendations: Portal

- Portal must be **continuously improved** based on **feedback**
  - **User engagement** is the **key** to building the CDS

- User interface must be **customisable** by user
- Provide **different views** to different users
  - Depending on their level of expertise and domain knowledge

- It should be possible to **browse** the content of the CDS **without login**
- Any **registration** requirement should be as **simple** as possible
- **Login** will be **required** to get **access** to actual data, products and services
  - Access statistics can be collected, for **reporting** and **capacity planning**

Findings and recommendations: User community

- User **forum**
- **Training** facility

- User should **learn by example**
  - A series of **use cases** should be presented

- Web based **help desk** for support

- "**Find an expert**" facility must be provided
  - A **knowledgeable** source on how to **interpret** data and products from the CDS

Findings and recommendations: Presentation

- Graphical data must be presented to the users in a **consistent** manner, to ensure a **unified** “look and feel”

- Presentation of information about **uncertainties** to **non-expert** users will need special consideration.

Findings and recommendations: Data and products

- Data in the CDS will primarily be in **binary** form
  - Support for **text documents** provided they are supplied with adequate **metadata**

- **No conclusion** on whether or not **socio-economic data** should be hosted by the CDS
- Limited to provision of URL

Findings and recommendations: Metadata

- S**uppliers** to the CDS will have to follow agreed **data management principles**
- Provision of detailed and accurate **metadata** information
- All data and products should be referenced by a **Digital Object Identifier** (DOI)

Findings and recommendations: Data Policy

- All content of the CDS should be freely available **without restriction** (Open Data)

- Support for commercial data and products will be considered at a **later stage**.

Findings and recommendations: Toolbox

- Ability to **visualise** them data product in the data portal

- Ability to download data **must include** facilities for:
  - **Sub-setting** large datasets,
  - **Re-gridding**/re-projecting gridded products
  - Performing f**ormat conversions**

Findings and recommendations: Toolbox

- More **advanced** tools (the "**Toolbox**") will be provided:
  - Users will be able to **download** these tools

  …or…

  - **Invoke** them from the **data portal**:
    - A list of predefined **workflows** that can be **parameterised**
    - Workflows operate **directly** on the data and products in the CDS
  - Toolbox should **preserve quality information** associated with the **input data**
  - Results of the workflows should be available for **visualisation** in the portal

- Tools could be based on CDO, NCO, Metview, etc.

Findings and recommendations: Toolbox (cont.)

- Providing users with the possibility to **upload** their **own data** as input of these tools **raises** several **issues** that will have to be considered carefully

- Similarly, whether or not the **results** of these workflows can become **part of the CDS** should be reviewed at a **later stage**

Findings and recommendations: Quality of Service

- An expected **large number of users** will be using the system **simultaneously**
- No one should accidentally (or maliciously) bring the system **down** by submitting **unreasonable** requests
- Workflows will run under a **scheduler** with controlled **quality of service** based on **queues**, **limits** and **priorities**
- **Caching** of results will also contribute to the performance of the toolbox

Findings and recommendations: Machine to machine access

- A **web-based API** should be provided to allow **bulk downloads** and **scripted access** to the CDS.

- The **R** and **Python** programming are good candidates to interact programmatically with the CDS

Findings and recommendations: Cloud Computing

- **Cloud** infrastructure (private, public or hybrid) must be **envisaged** for running the workflows

Findings and recommendations: Interoperability

- Other Copernicus services
- INSPIRE
- GEOSS
- WIS
- GFCS
- WCRP

Findings and recommendations: Standards

- The use of standards is **essential** to ensure interoperability between the CDS and its **suppliers**, as well as between the CDS and its **users**
- Standards:
  - ISO (ISO-19115)
  - OGC (WMS, WPS, …)
  - WMO (GRIB, BUFR)
  - Unidata (NetCDF, CF)

- C3S should **keep a close link** to the relevant **standardisation committees**
- SIS users and providers may also be able to provide advice on issues of governance and **relevant standards** from their own specific domains

Findings and recommendations: Monitoring

- The CDS is **continually monitored** and **assessed**

- A number of **key performance indicators** (KPIs) must be defined for this purpose.
  - E.g. year-on-year increase in the number of active users

Findings and recommendations: Implementation

- Initial development of the CDS is to **start** with a set of **basic functionalities**
  - Allowing early users to provide **feedback…**
  - …and then implement **more advanced features** over time

- ECMWF must work in **close collaboration** with competitively selected **contractors** to implement the CDS in an **iterative fashion**
  - An **agile development methodology** is preferred

TENDER 1

- A **distributed data store** that will be built upon **existing infrastructures** available at each of the suppliers.
- A **toolbox** that will contain software components that can be used to perform **computations** on the content of the data store, in a **distributed** fashion, under strictly controlled **quality of service** constraints;
- A **centralised catalogue** that will describe the holdings of the distributed data store, as well as the **tools available** in the toolbox;
- A **broker** that will forward data and services requests to the relevant suppliers;
- A **web portal** that will allow users to discover and interact with the content of the climate data store. This will enable users to **browse** and **search** the catalogue, **submit** data requests and perform **computations** on these data. The portal will also act as a **content management system**, in order to provide **documentation**, **training** material, **support** pages, etc.
- Consider use of **Cloud Computing** for the computations

TENDER 2:

- **Tools** that perform **basic operations** on data (e.g. statistics, values at points,…)
- **Workflows** that **combine** tools by chaining them
- **Applications** that make use of **workflows** and **selected data and products** of the CDS, to build **interactive web-page** allowing end-users to interact with the CDS
- Users will also be able to **share** with other users applications they have developed
- Study how workflows can be **orchestrated** so that computation to be performed **close to the data** and **data transfers** to be **minimised**
- Implement a series of **reference applications** that will demonstrate all functions offered by the toolbox
- Study **“Big Data”** techniques, such as map-reduce, can be applied to the CDS, and what are the implications on the infrastructure

Questions?