



ILMATIETEEN LAITOS  
METEOROLOGISKA INSTITUTET  
FINNISH METEOROLOGICAL INSTITUTE

# Tomorrow's Data Discovery and Availability Services at FMI

**Ilkka Rinne**

**11th Workshop on Meteorological Operational Systems**

**13<sup>th</sup> November 2007, Reading, UK**



# Outline

- **Why are data catalogues and metadata important right now at FMI?**
- **Service oriented data availability**
  - A typical scenario
- **Quick look to the standard work for data catalogues and meteorological metadata**
- **Internal data catalogue implementation issues**
- **Getting hold of the actual data versus metadata**



## Some Current Development Projects at FMI

- **Gradual redesign and solidification of the whole operational system architecture as part of the internal SISU project is currently (Nov. 2007) in the design phase.**
- **National information interchange interfaces and externally visible data and/or service catalogue service + other SOA style technologies are prototyped in a research project in co-operation with Vaisala during 2008.**
- **Preparations for publishing metadata and building standard based, interoperable data catalogues (for the EU INSPIRE directive, but also other national agencies)**



# Data Availability Services?

- **How to find the data that *should* be stored somewhere?**
  - Or know for sure, that we don't have it?
    - If so, how to get it from somewhere else on demand?
- **How to define and gather the necessary metadata for all the incoming and self-produced meteorological information?**
  - Standard metadata description languages for meteorological data would be nice to make national and international information interchange easier
- **How to establish a minimal but sufficient set of data transfer methods and data formats?**



# Data Availability Knowledge Is Essential

- **Need to take full advantage of the improving resolution of the numerical forecast and nowcasting data in the highly customized end-user products**
  - No time to generate products that are not needed right now. Instead do more on-demand processing
- **For on-demand production to be efficient and produce up-to-date products, good and fast data availability services are needed.**
- **The machines alone cannot do the forecasting well enough: The meteorologists make operative decisions in selecting and correcting the data based on multiple data sources and their expertise.**
  - Need to find relevant comparison data efficiently



# Challenges In The Current Operational System

- **Different query interfaces to the different data storages**
  - SQL, C++ API, direct file-system browsing, own in-house query languages,...
- **For some data, scanning file system directories is the only way to find out if the data for certain time is available or not.**
- **The same data might be scattered around the system in different formats.**
- **No followable audit trails for the production chains from the raw data arrival to the delivered customer products.**

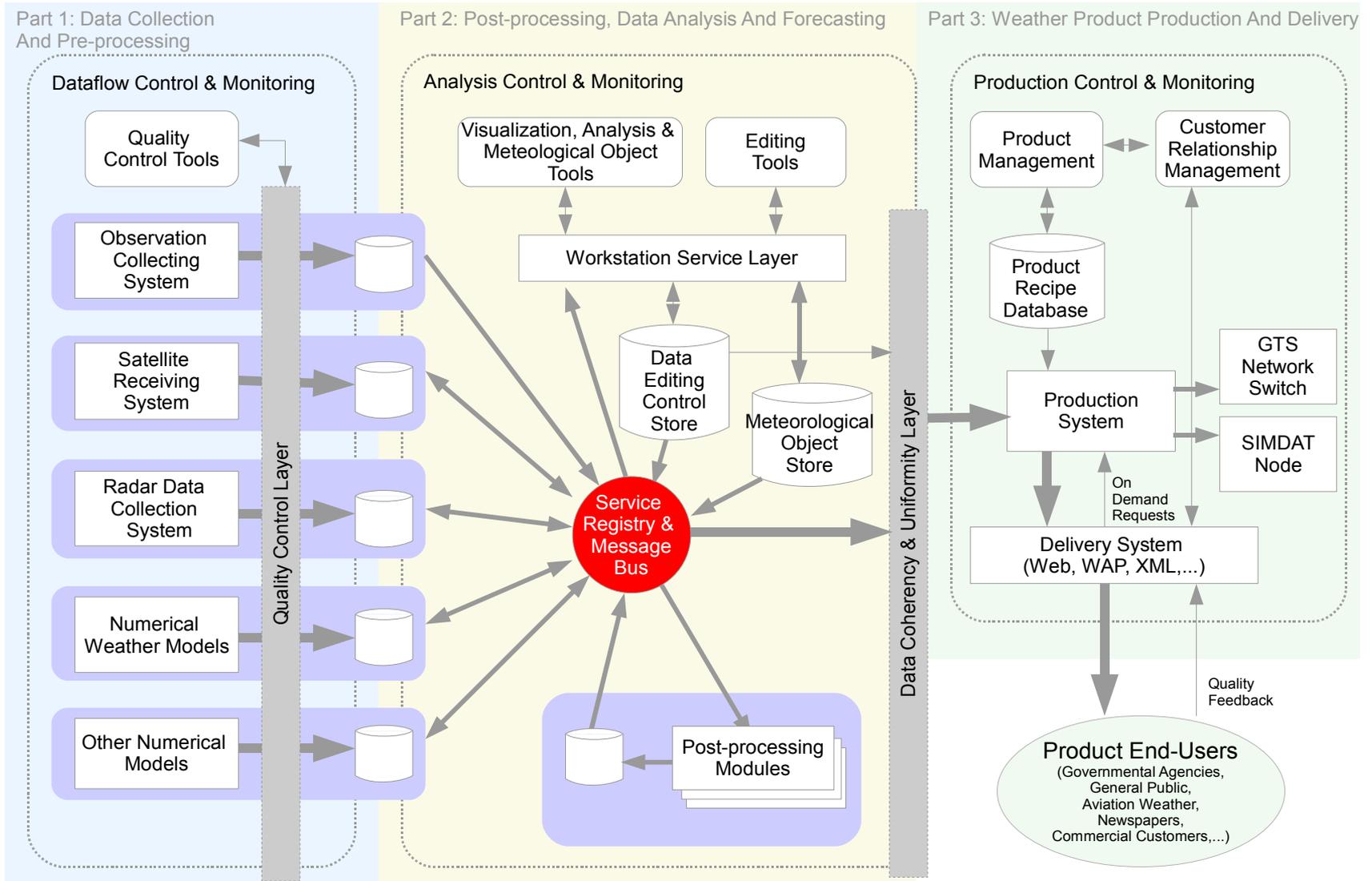


# Use Service Oriented Architecture Internally

- **Many different internal data providers, but one central Service Registry**
  - What kind of data is offered by which provider?
  - How can the queries for available metadata be filtered upon a query?
  - In which formats and packages can the data be retrieved?
- **Each data provider is responsible for**
  - storing the data,
  - gathering the necessary metadata, and
  - providing the query interface for the available data.

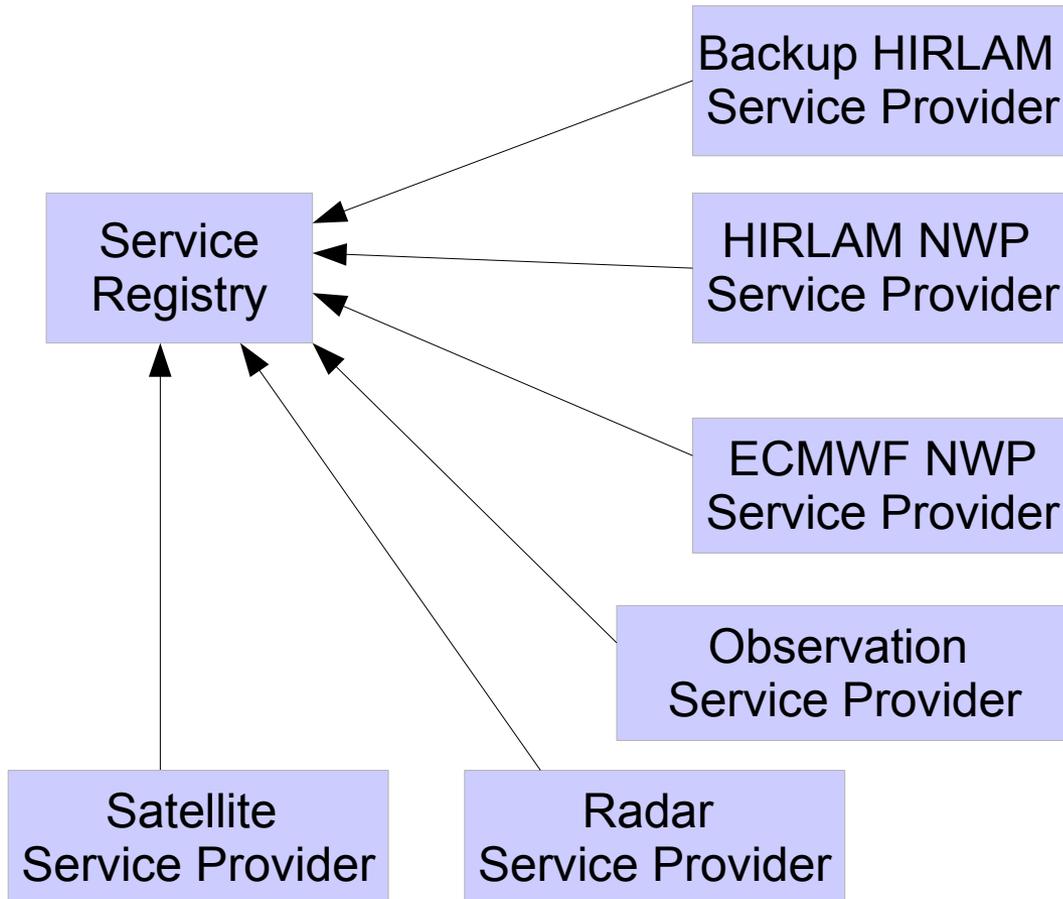


### Planned FMI Weather Information System, Level 0





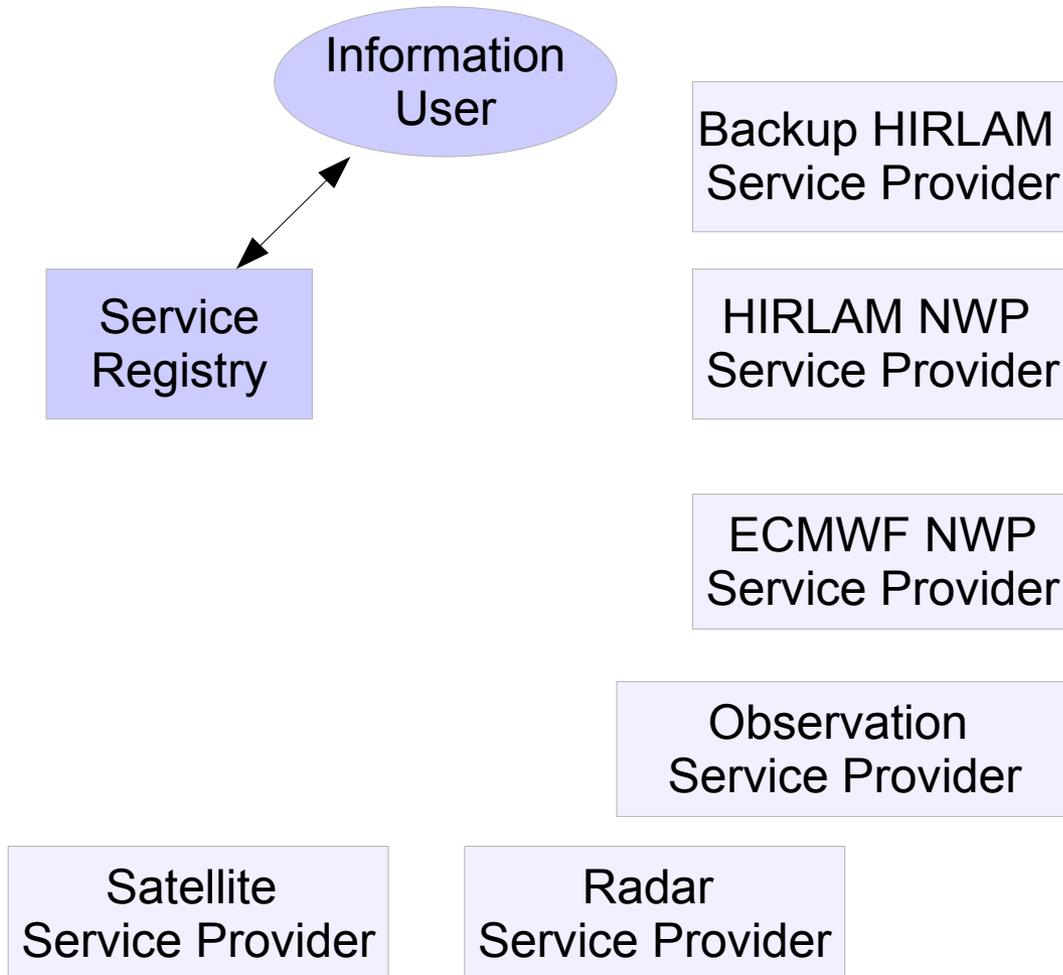
# The “SOA Inside” Scenario



Data & Service Providers register their services to the registry when becoming on-line and going off-line.



# The “SOA Inside” Scenario

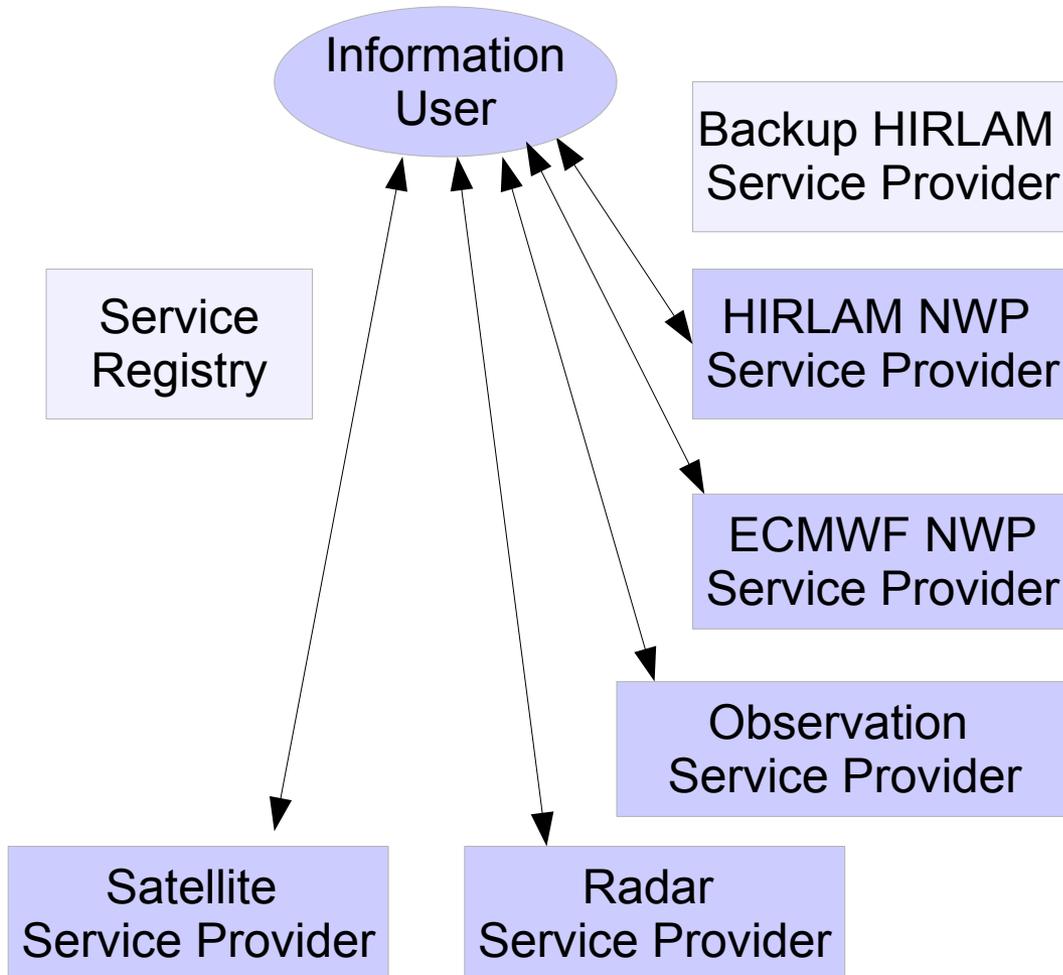


Information user **requests services** from the Service Registry based on requested data types, requested spatial boundaries and acceptable data formats.

The returned data contains **connecting information** and allowed **query parameters** for each matching service. Records contain an **expiration time** after which they may no longer be used.



# The “SOA Inside” Scenario

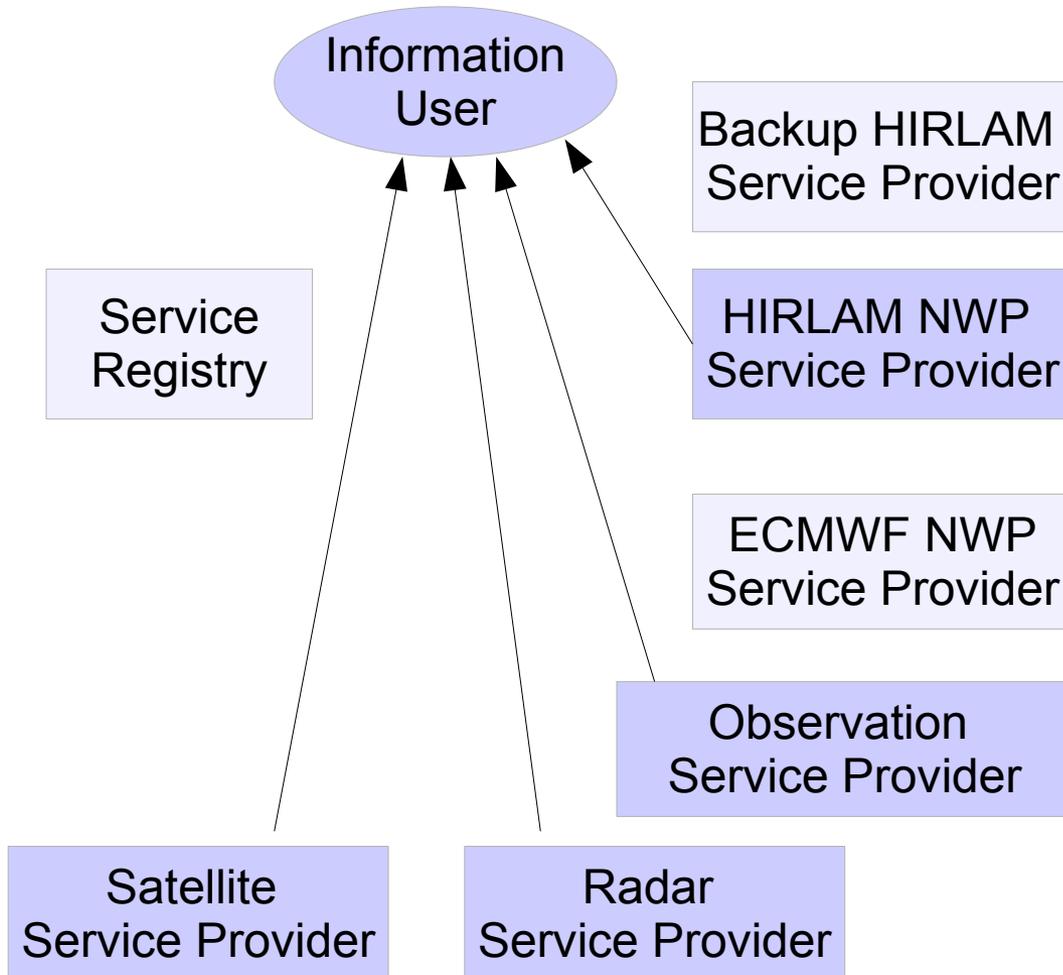


Information User makes **requests** for currently **available data** to the Service Providers based on the information returned from the Service Registry. Time-based query criteria is typically used at this point.

**Metadata records for matching data** currently available are returned from each Service Provider.



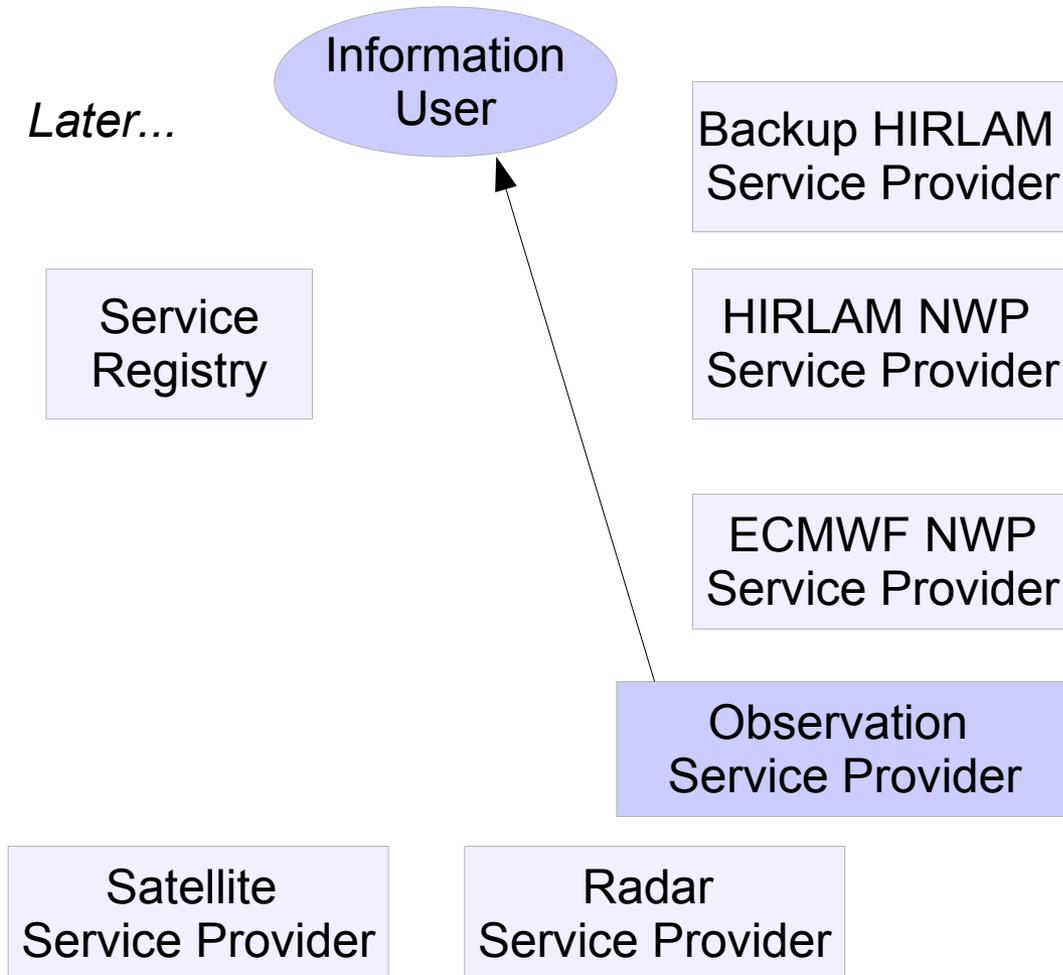
# The “SOA Inside” Scenario



Information User selects the **data to download** based on the replies from the Service Providers, and retrieves the actual data using a commonly agreed data transfer methods and data formats for each service.



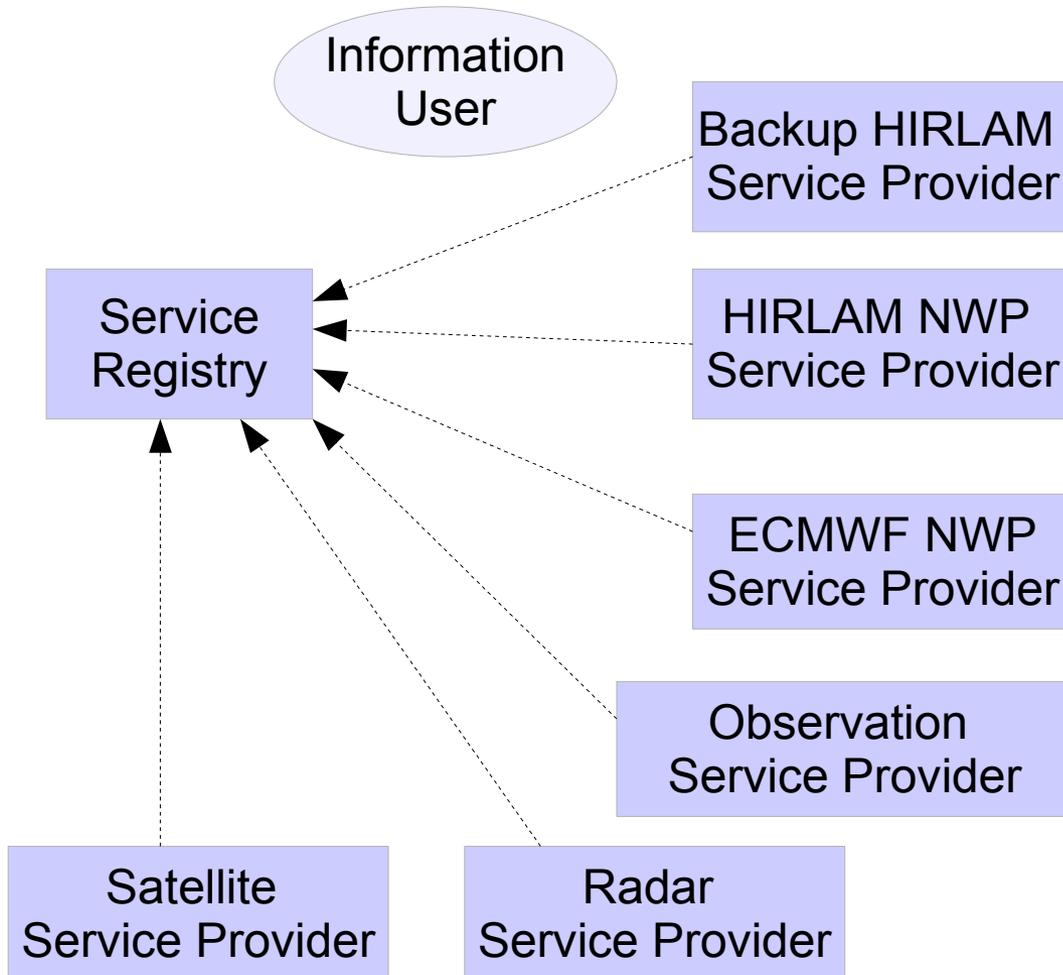
# The “SOA Inside” Scenario



Information User may use a previously retrieved service metadata to make further queries directly to the Service Providers until the service metadata expires.



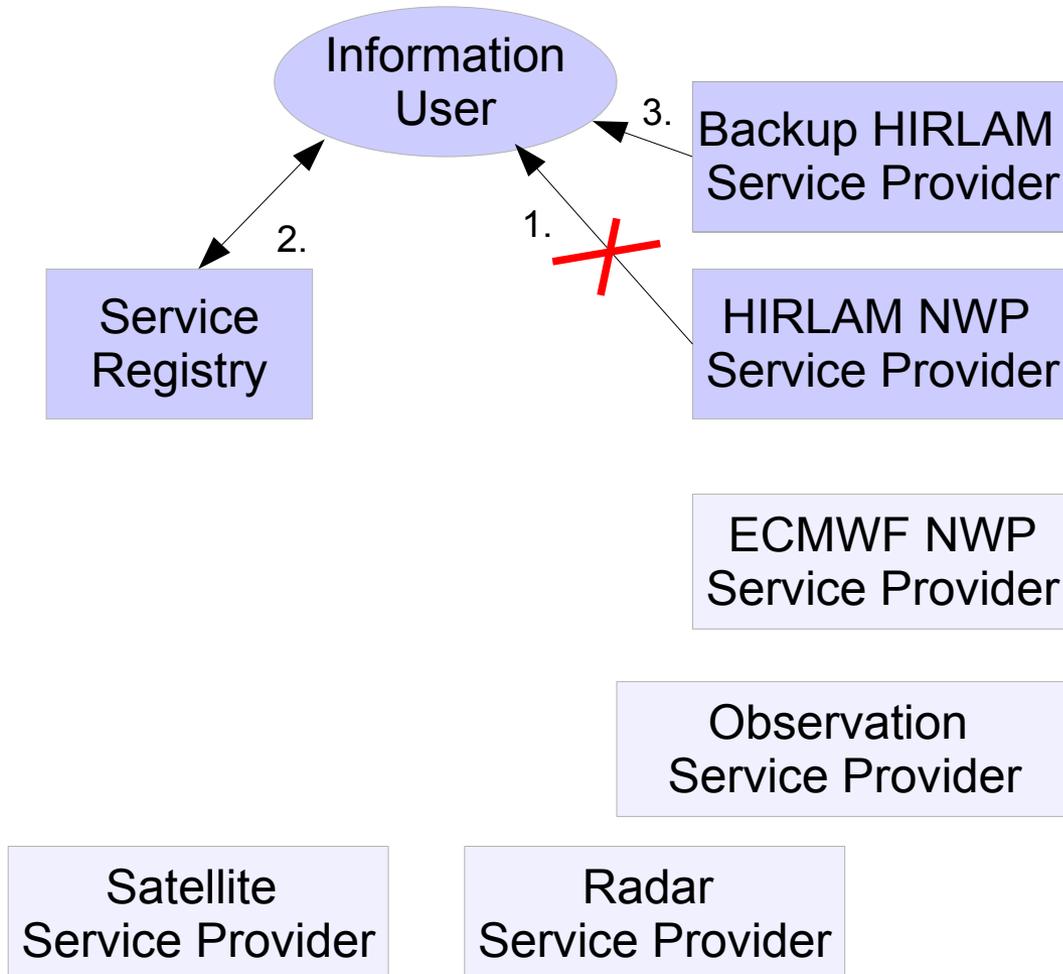
# The “SOA Inside” Scenario



Service Providers send “**heartbeat**” messages to the registry to notify it about their **status and load statistics**. The registry updates its service catalogue and load-balancing information according to this information.



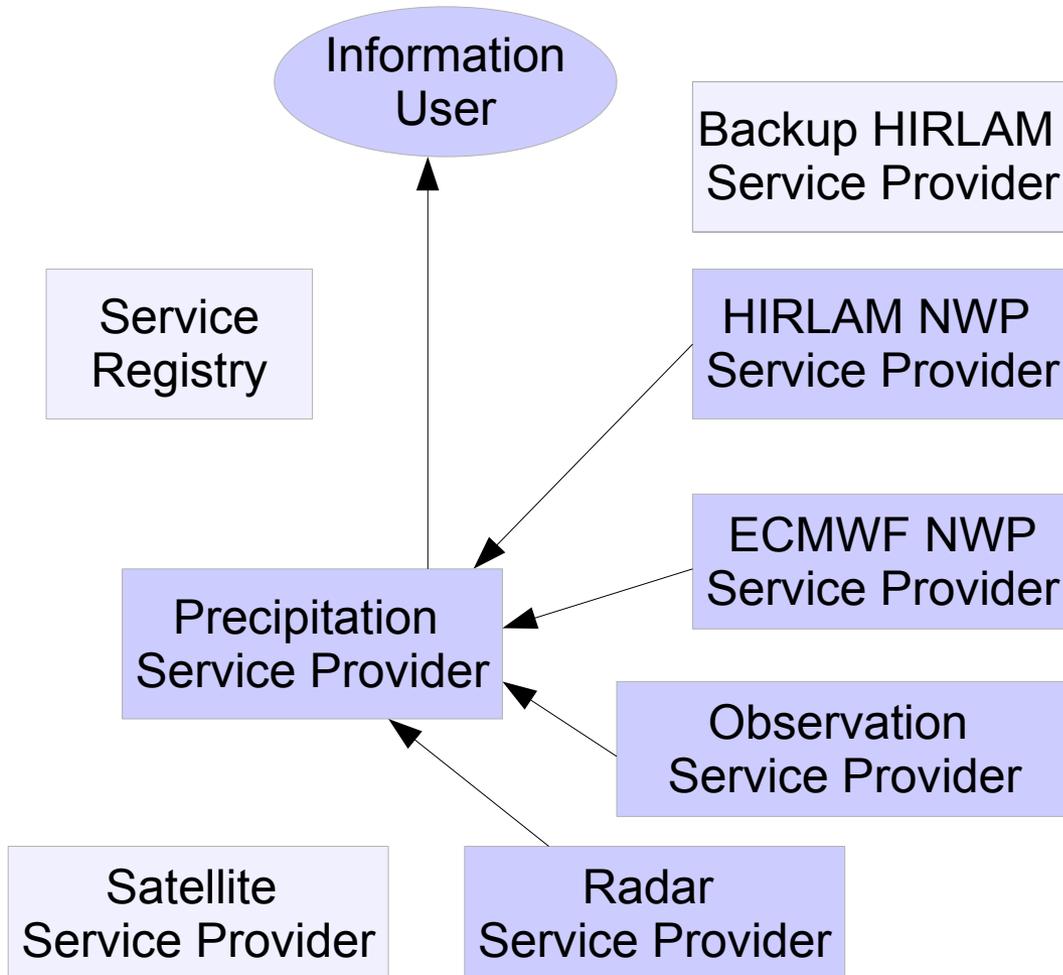
# The “SOA Inside” Scenario



If communication problems should occur with the used Service Providers, the Information User reports this to the Service Registry and requests information about other providers for the same type of information.



# The “SOA Inside” Scenario



A post-processing systems act both as Information Users for the required input data and Service Providers for their specific output data.



# “Standard” Catalogues And Registries

- **Generally a catalogue is an open and well defined collection of GML application schemas, phenomena directories, controlled vocabularies, service bindings etc.**
  - A common place for discovering data-centric services and their communication interfaces
  - Registry = governed catalogue
  - A key enabler in data and services interoperability
- **OpenGIS Catalogue Services Specification**
  - Abstract and generic
  - HTTP binding = Catalogue Service for Web (CSW)
  - IIOP/CORBA binding is also specified



# “Standard” Catalogues And Registries

- **ebXML Registry Information Model (ebRIM) used as registry model for the majority of OGC Catalogue Service implementers**
  - Specifies how the metadata is organized in the registry
  - OGC prefers this registry information model over for example a more general UDDI
- **Question: Is the CSW/ebRIM fast & reliable enough (and not too complex) for operational use?**
  - 100 msecs is a much longer time within the operational production system than in web browsing.

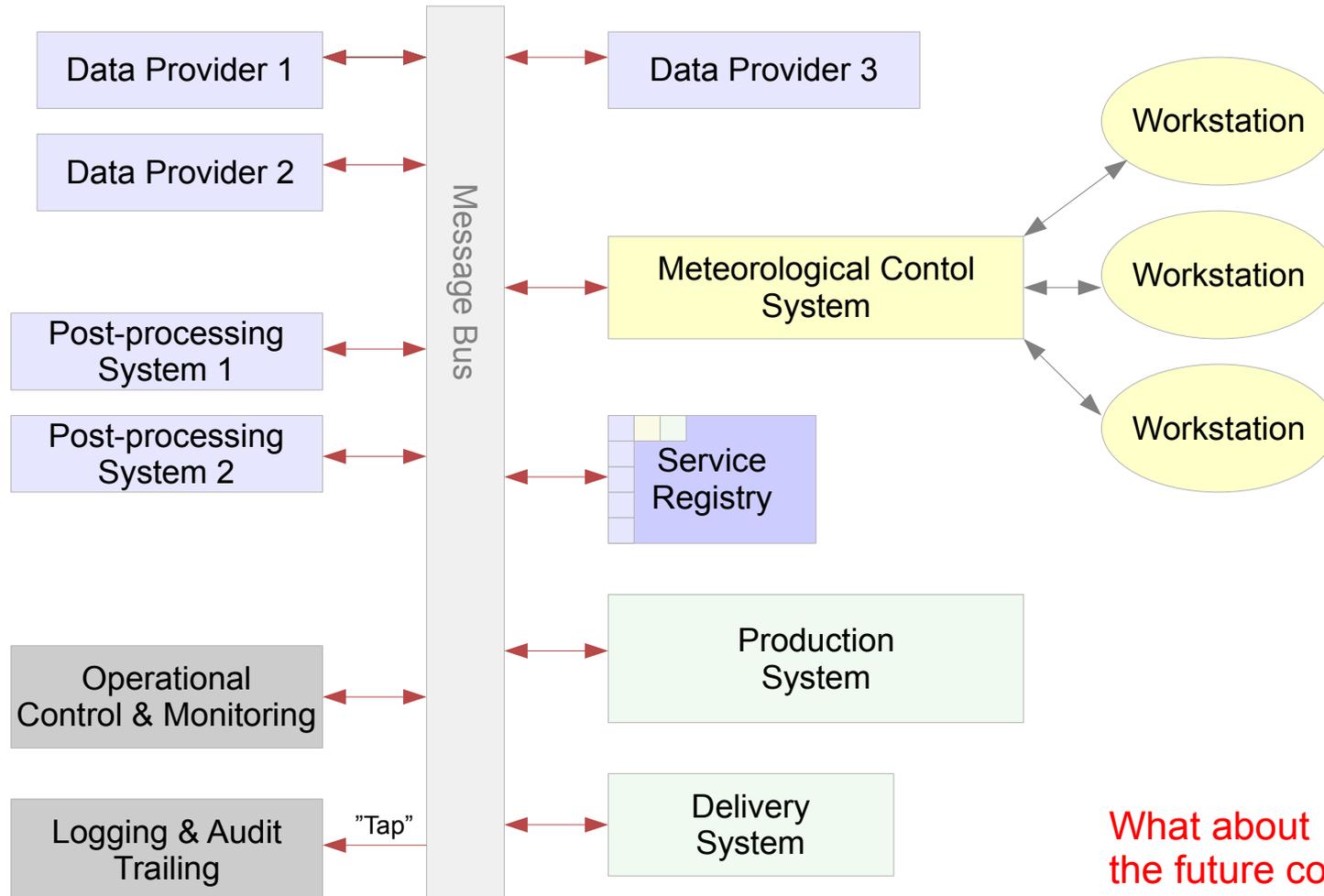


# Internal Catalogue Service Implementation

- **Use the “standard” catalogue information model and metadata languages**
  - Provide a common language for international information interchange
  - No use inventing the wheel once again
- **But the implementation of the internal catalogue service interface need not to a web service (that is HTTP) based:**
  - Message Orientation would be interesting: asynchronous data availability queries, status notifications and loose coupling between the system components would introduce a lot of flexibility
  - Better quality of service and less overhead than HTTP



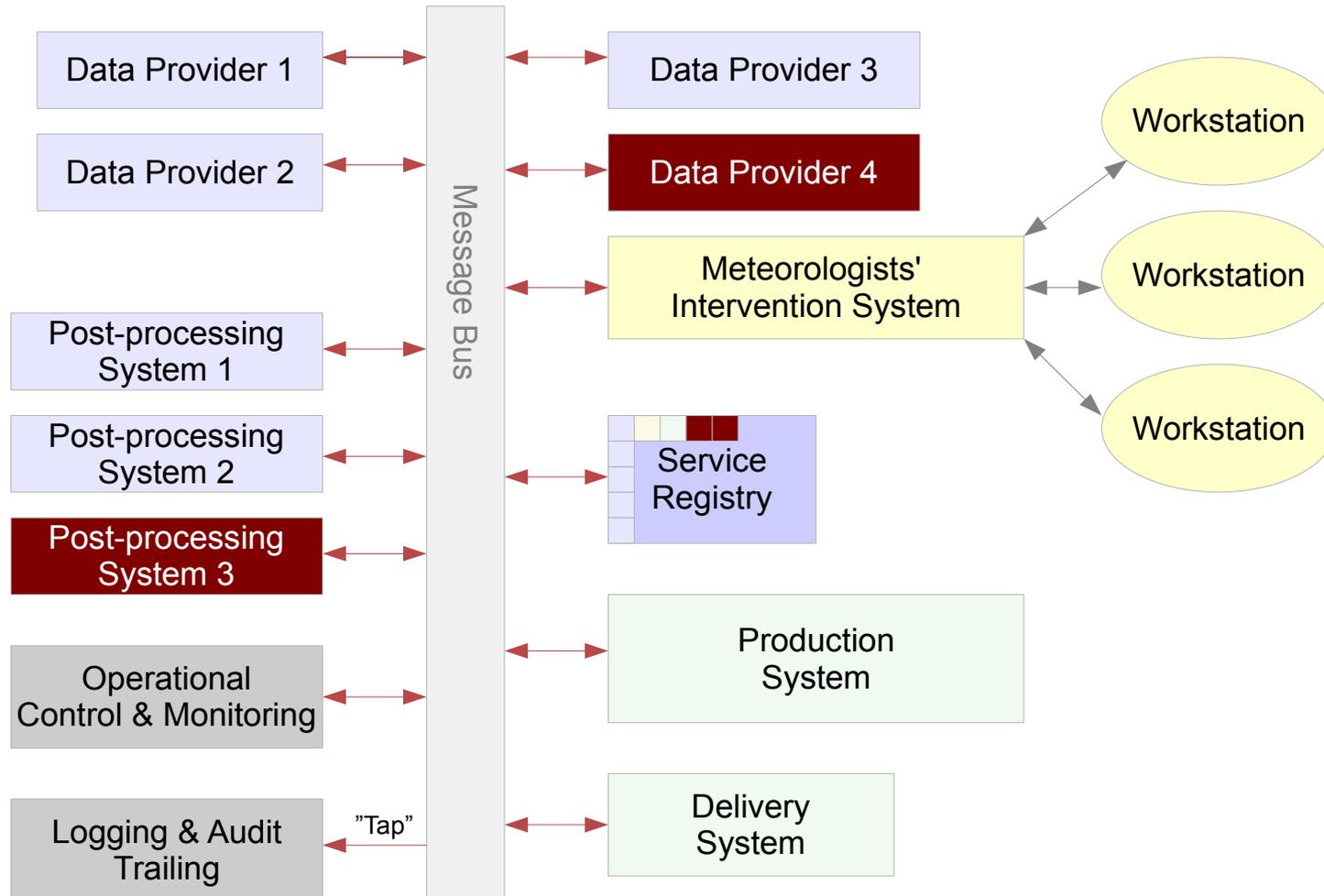
# Message Oriented Data Availability?



What about  
the future components?



# Message Oriented Data Availability?





# Data Transfer

- **A Java-based Messaging System (JMS) for the availability messages, notifications, heartbeats,...**
  - with adapters to systems written in other languages
- **Transferring large datasets over a common messaging system is probably not a good idea**
  - Relatively small messages make things much easier for the message bus, which needs to be as simple and reliable as possible.
- **Why copy the bulky data files over the network if we could read them directly using a shared file system?**
- **If the physical data transfer is needed, use existing efficient data transfer mechanisms**
  - FTP, rsync, you name it.



# Conclusions

- **The importance of efficient data availability services will grow rapidly in the near future.**
- **The need for standardized meteorological metadata, data catalogue service interfaces and catalogue implementations is obvious**
- **Adhering to the standard repository information models and data description languages also internally makes the future integration easier**
- **Standing on the shoulders of giants: We should bring all the expertise used in making the standards into a proper use.**
- **Message orientation could provide a solid backbone for the next generation met. operational system**