D4SCIENCE INFRASTRUCTURE

A FACET-BASED OPEN AND EXTENSIBLE RESOURCE MODEL FOR RESEARCH DATA INFRASTRUCTURES

Leonardo Candela, **Luca Frosini**, Pasquale Pagano

ISTI – CNR (Pisa, Italy)

The RDI landscape

Research Data Infrastructure (RDI)

harnesses the accumulating data and knowledge produced by the communities of research, optimizing the data movement across scientific disciplines, enabling large increases in multi- and inter-disciplinary science while reducing duplication of effort and resources, and integrating research data with published literature

- An RDI is intrinsically a system of systems federating data, tools, services, and actors belonging to heterogeneous scientific communities
- An RDI is disassembled into its component systems, each organization and managing their own resources
- An RDI is naturally geographically distributed
- An RDI evolves accordingly to the evolution of the component systems

Challenges

To deal with **heterogeneity** with respect to:

- Diverse set of managed resources
- Diverse model for describing resource characteristics
- Diverse workflows governing registration and update of resources

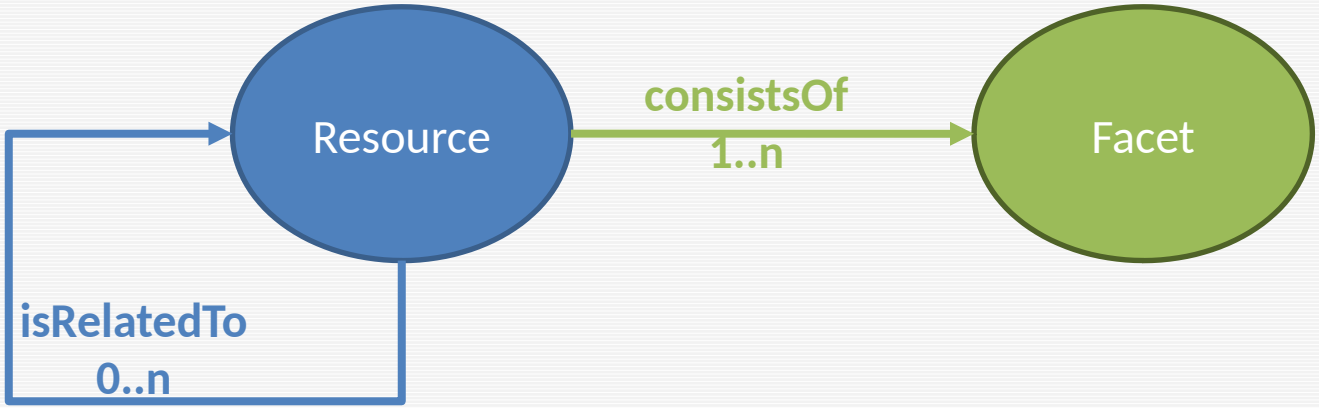
Goals

To define a **flexible** model capable of manage the heterogeneity

- Open-ended set of manageable resources
- Open-ended model for describing resources
- Ability to evolve with the evolving needs of the infrastructure at no cost for its clients
  - by supporting new resource types
  - by supporting evolution in the way a resource is described
  - by supporting the same resource type described by using different models

The Model

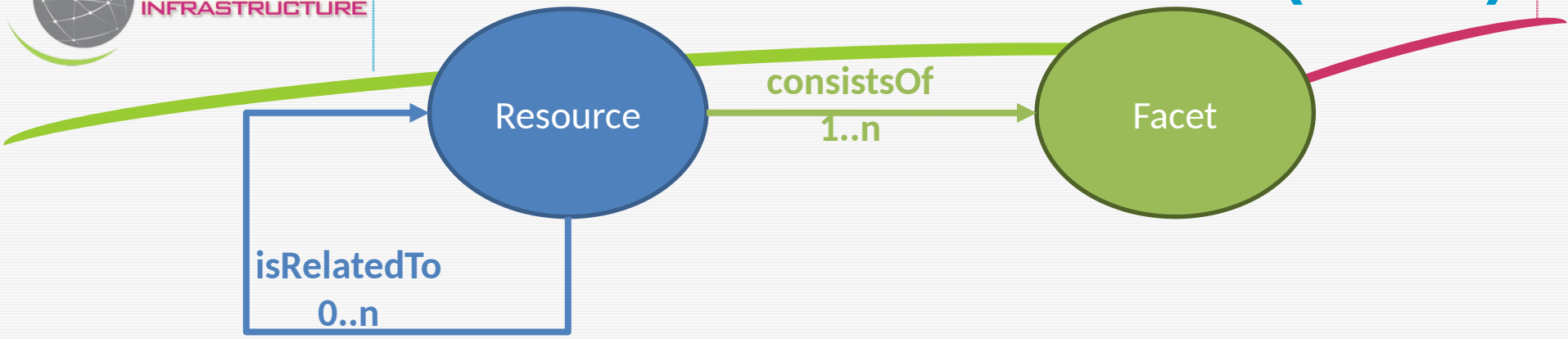
An apparently simple model to govern the RDI complexity



The Model (cont.)

- **Resources :** i.e. entities representing a *"thing"* to be tracked/ managed and are characterized by a number of Facets
- **Facets :** i.e. entities contributing to *"build"* a description of a Resource. Every Facet captures a certain aspect of the resource and is characterized by a number of properties
- **isRelatedTo :** i.e. a relation linking any two Resources
- **consistsOf :** i.e. a relation connecting each Resource with one of the Facets characterizing it

The Model (cont.)



- Any **Relation** has a direction, i.e. a **"source"** (out bound of the relation) and a **"target"** (in bound of the relation).
  - Any relation can be also navigated in the opposite direction
- It is not permitted to define a **Relation** having a **Facet** as ***"source"***. In other words:
  - It is not permitted to define a **Relation** connecting a **Facet** with another one
  - It is not permitted to define a **Relation** connecting a **Facet** with a **Resource** (as target)
- A **Facet** instance can be linked (by **consistsOf** or any specialization of it) from different **Resources**
- Every **Relation** is characterized by a number of properties

Entities and Relations Specializations

Starting from the simple and powerful model we defined a number of specializations to support the Research Data Infrastructure scenario

- We defined the type of resources
- We defined the relationships among those resources
- We defined the description of those resources through the definition of a number of facets

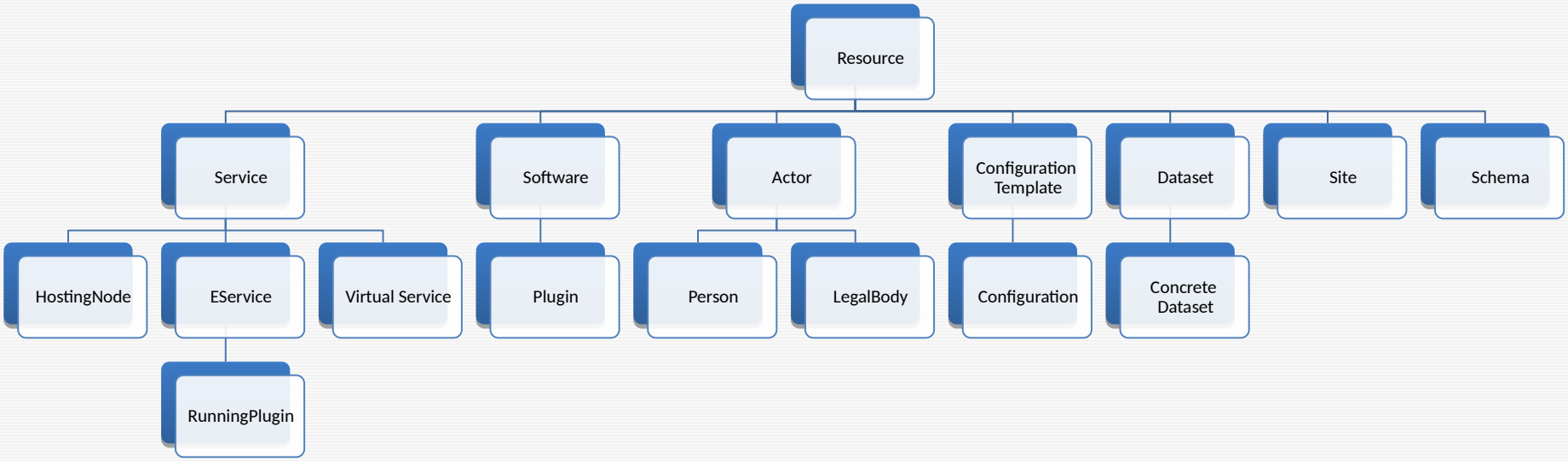
To manage the openness and the evolution of the RDI, the model can be easily specialized by defining

- new typologies of relations
- new typologies of entities (both resources and facets)

RDI : Identified Resources

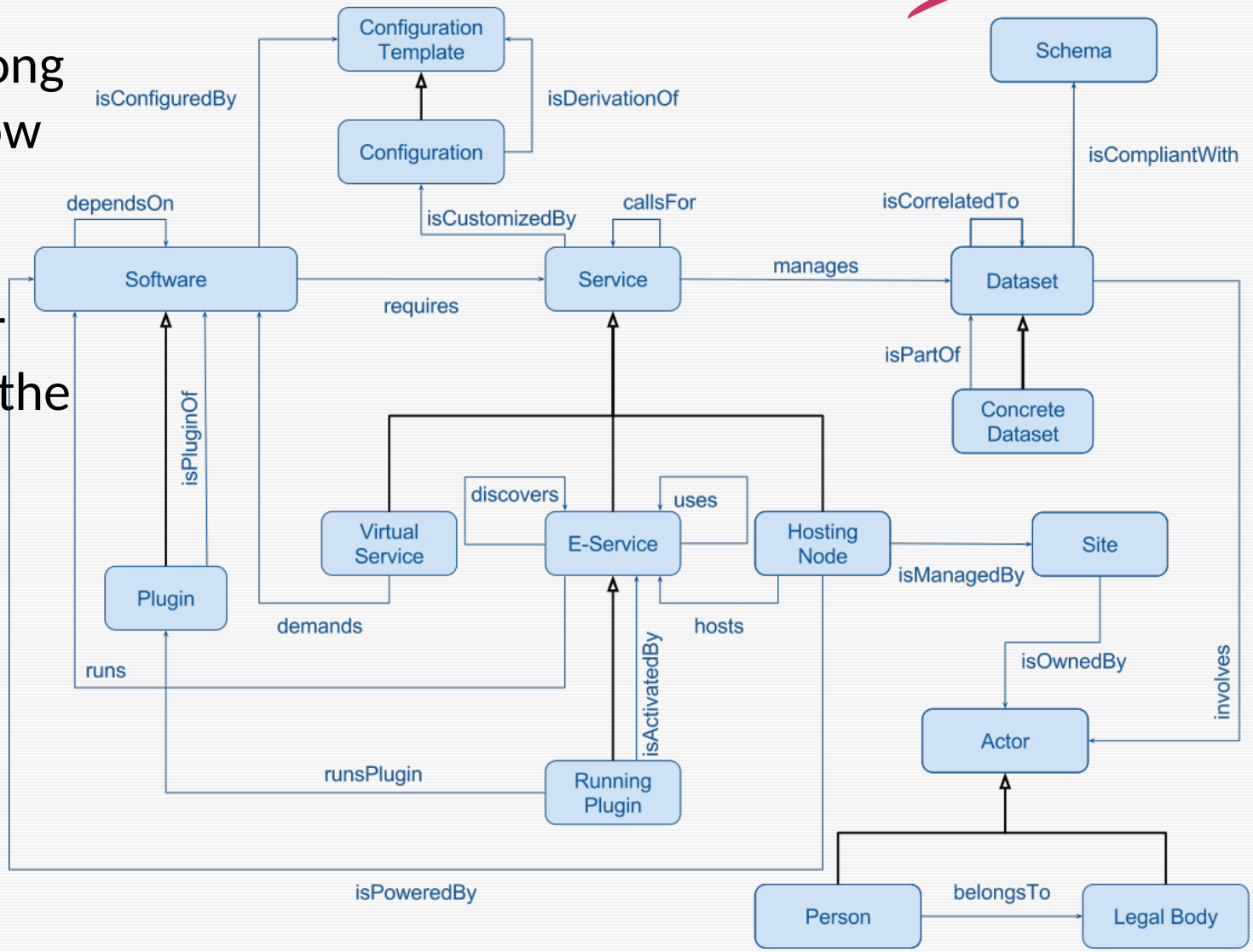
A Research Data Infrastructure

- manages Services, Software, and Datasets
- operated and provided by Actors
- provisioned by different Sites
- and configured to satisfy the needs of the Scientific Communities



RDI : Identified Relations

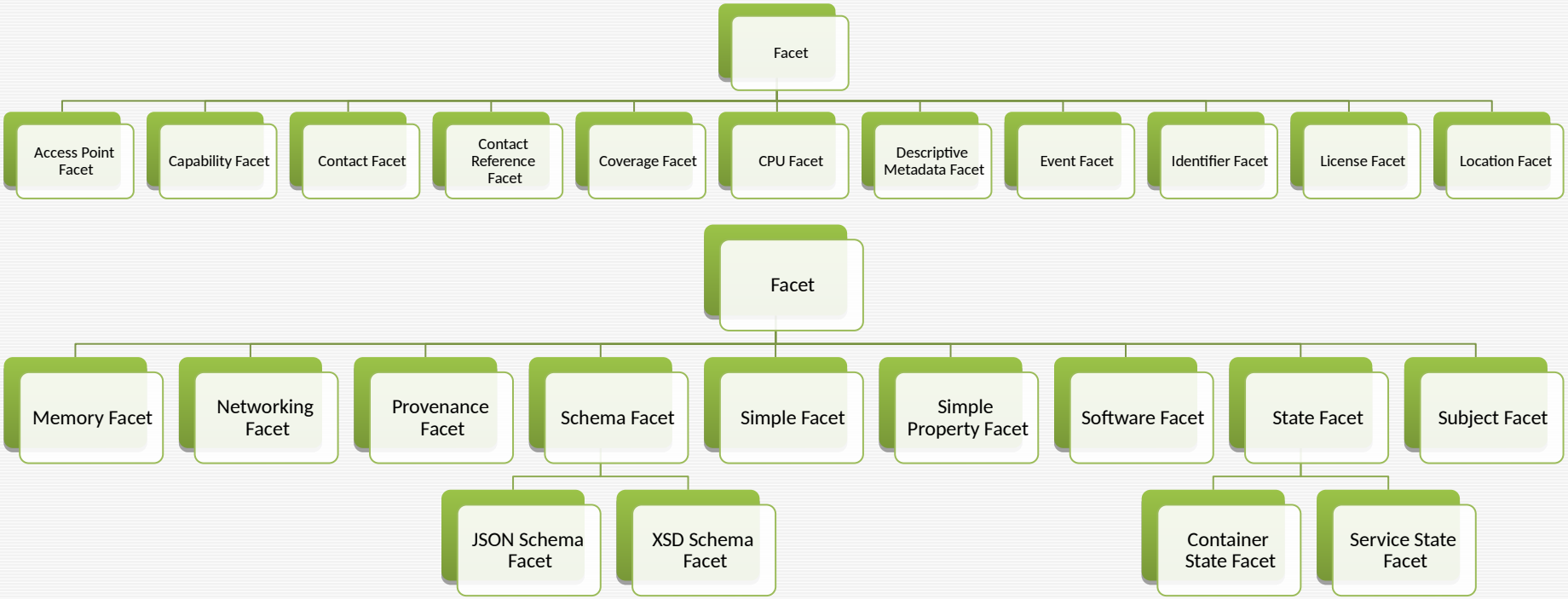
Relations among resources allow to define the semantic and the behaviour expressed by the relationship



RDI : Identified Facets

An RDI exploits software and services to manage the entire data lifecycle: discovery, access, harmonisation, publication, validation.

To support this scenario, a number of characterisations of the resources has been modelled through a number of Facets.



D4Science is a Research Data Infrastructure

- connecting **+4000 scientists** in **44 countries**
- Integrating data from **+50 heterogeneous providers**
- executing **+42,000 models & algorithms/month**
- provides access to over a **billion quality records** in repositories worldwide
- operating with **99,8% service availability**

D4Science hosts **+100 Virtual Research Environments** to serve the biological, ecological, environmental, and statistical communities world-wide.

CONTACTS

luca.frosini@isti.cnr.it

leonardo.candela@isti.cnr.it

pasquale.pagano@isti.cnr.it

Technical References

https://wiki.gcube-system.org/gcube/Facet_Based_Resource_Model