The LIDER Reference Architecture

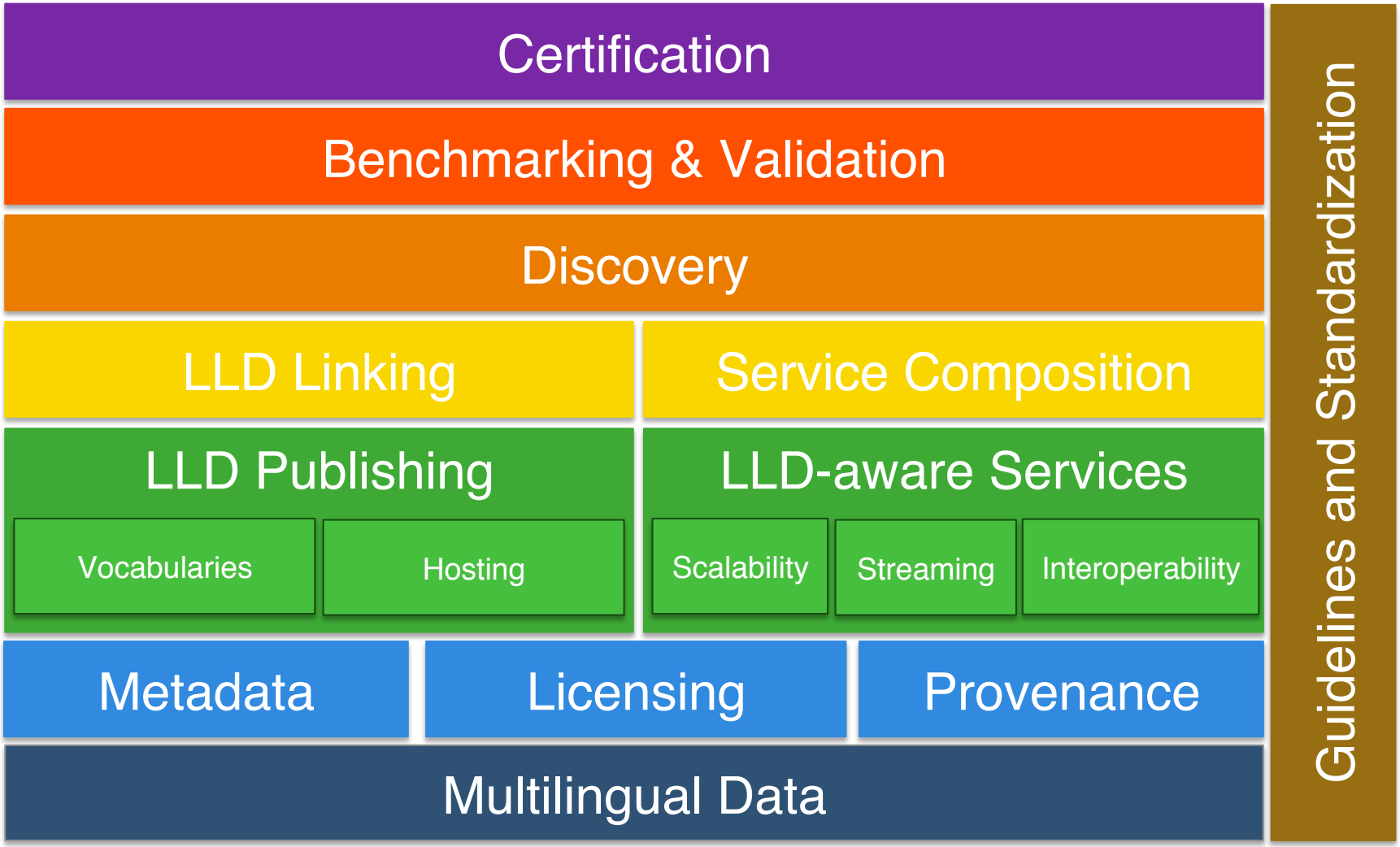
Philipp Cimiano
(representing the LIDER Project)
LD4LT Teleconference
March 5th, 2015

Goal

- **Goal:** Develop a Reference model that supports an ecosystem of linguistic linked data and the development of content analytics services on top of this ecosystem.

- **Key features:**
  - **Linked Data:** connected ecosystem of data and services, interoperability, supporting access by both humans and machines
  - **Semantic Technologies:** open web standards (OWL, RDF) for data description, SPARQL and HTTP as Web APIs
  - **De-centralization:** Web architecture, no central point of failure, no vendor lock-in, open standards

Reference Architecture



Reference Architecture

- Terminologies
- (Multimodal) Corpora
- Bilingual Dictionaries
- Parallel Data
- Translation Memories
- Ontologies
- Glossaries, Classification Schemas

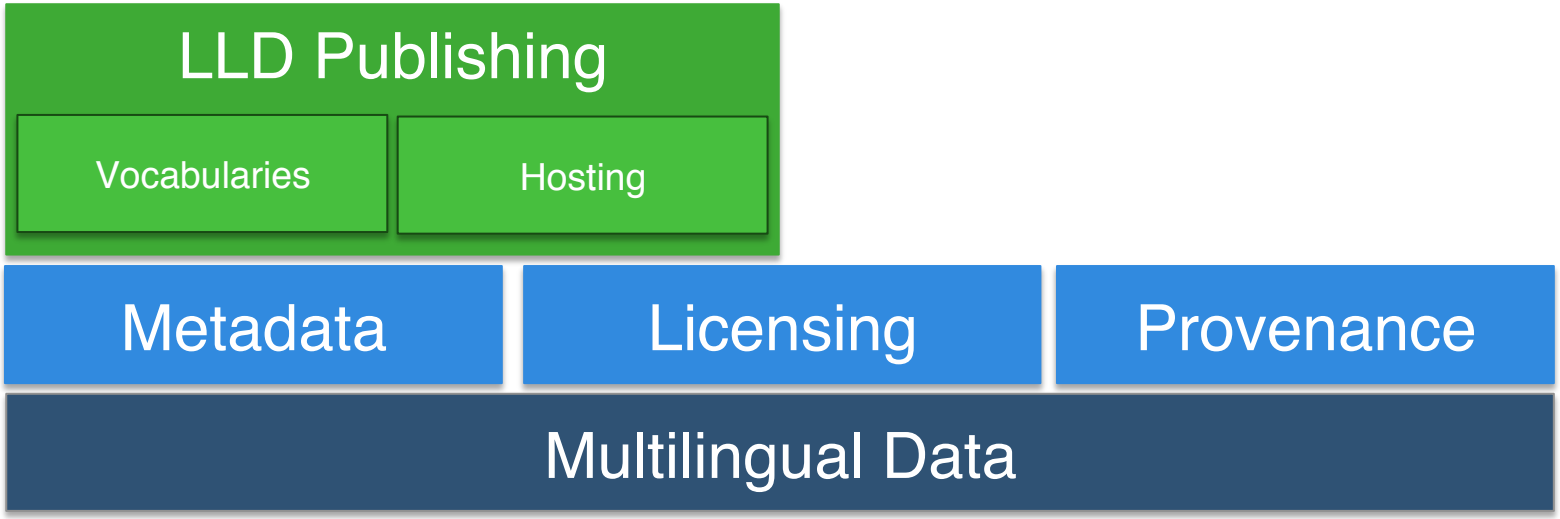
Multilingual Data

Reference Architecture

- **Metadata:** providing basic information about the dataset (author, language, structure), etc.
- **Licensing:** specifying the terms and conditions of use
- **Provenance:** describing the origin and processing history of data

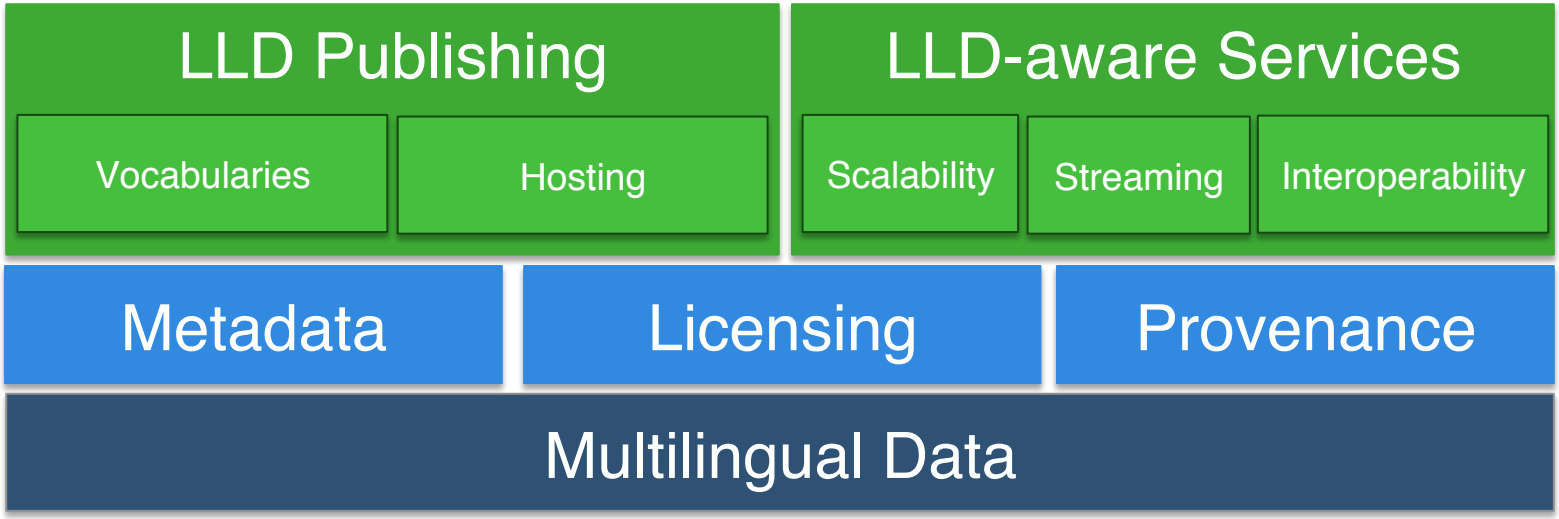
Reference Architecture

best practices, standards and tools for **publication** and **hosting** of LDL, **and vocabularies** for description and transformation of different types of resources (lexica, corpora, terminologies, lexico-semantic resources) into RDF/LDL Linguistic Linked Data (LDL)



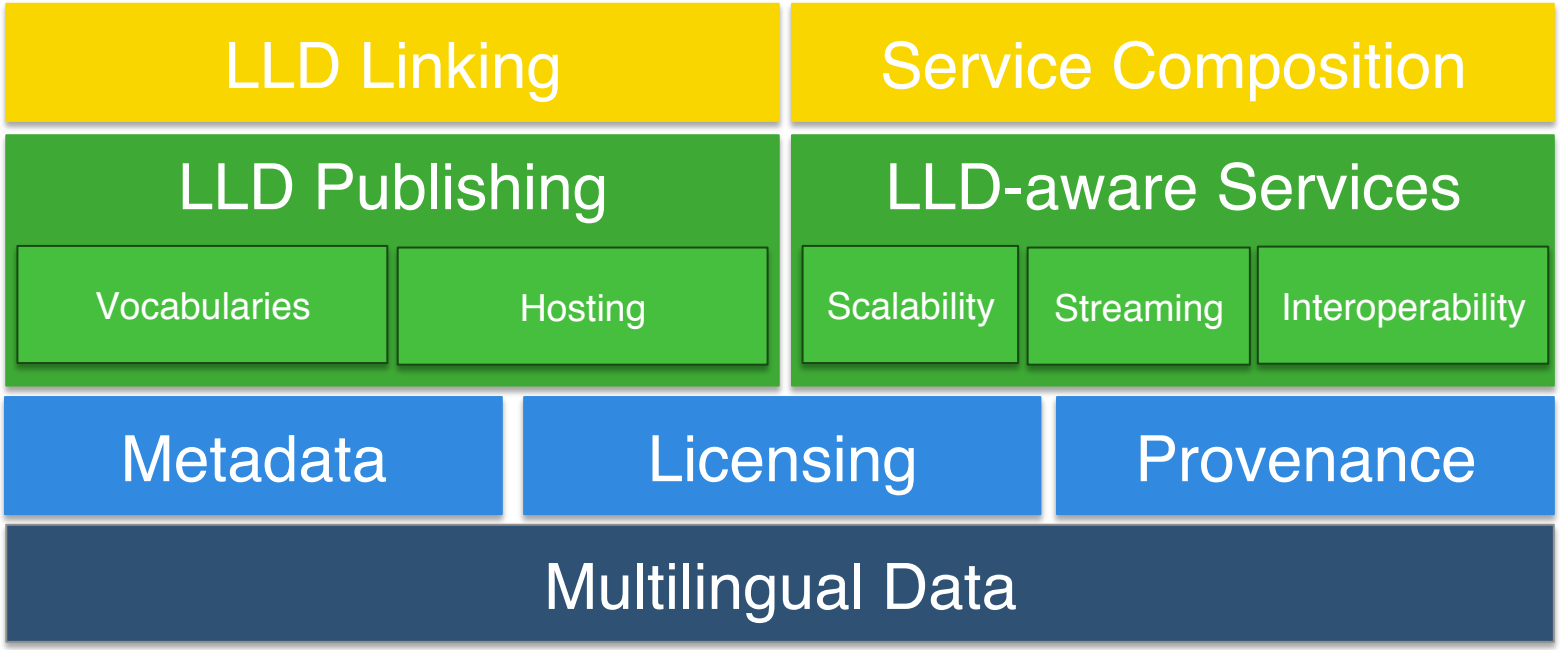
Reference Architecture

- **Scalability:** caching and non-centralized processing
- **Streaming:** process data in a stream fashion, thus reducing overhead of creating and closing connections
- **Interoperability:** common vocabulary to describe inputs and output of services



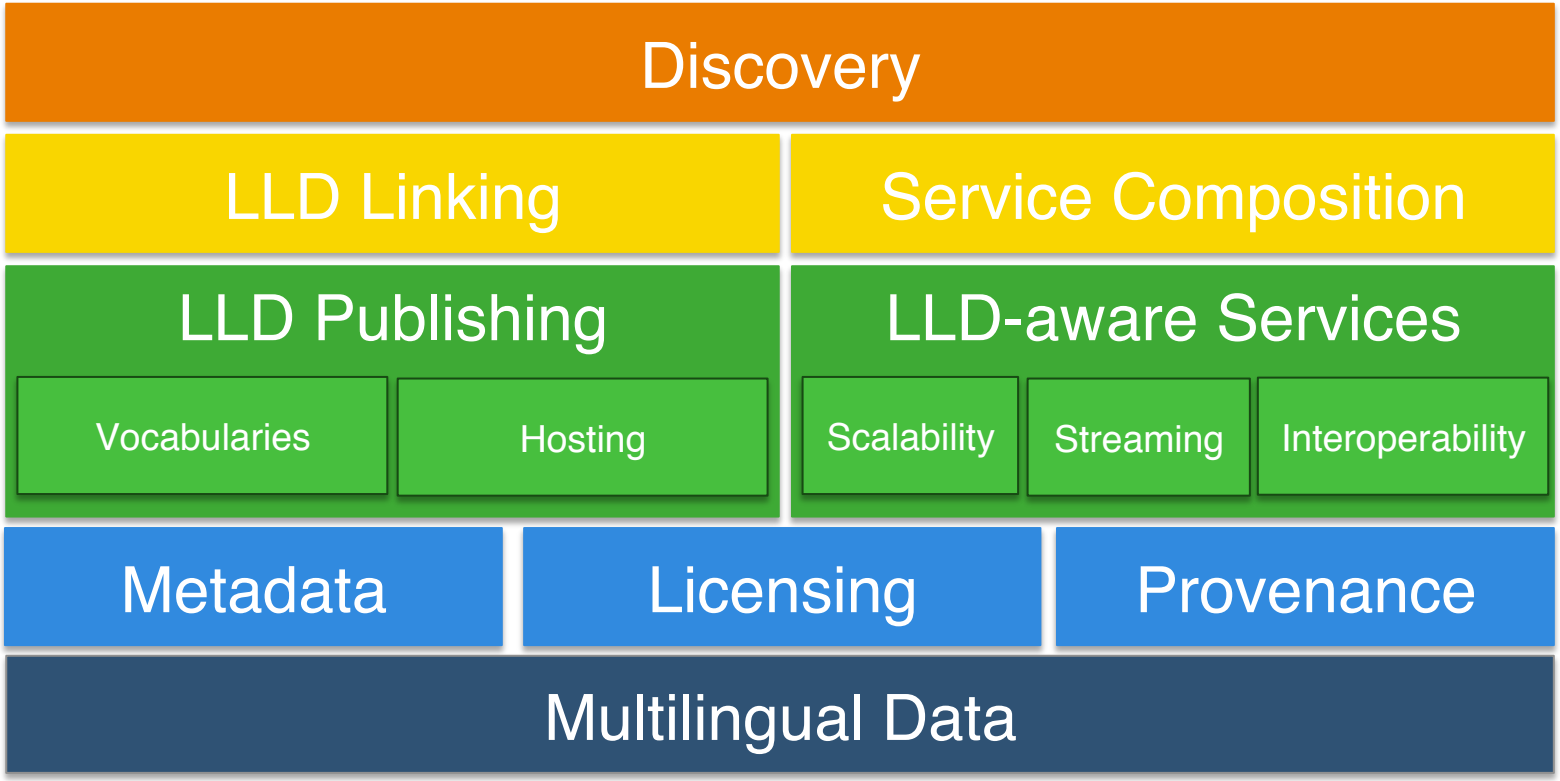
Reference Architecture

- best practices to supporting **linking of resources**, combination of data with different terms and conditions of use, in particular open and closed data
- support **composition of services** into complex workflows



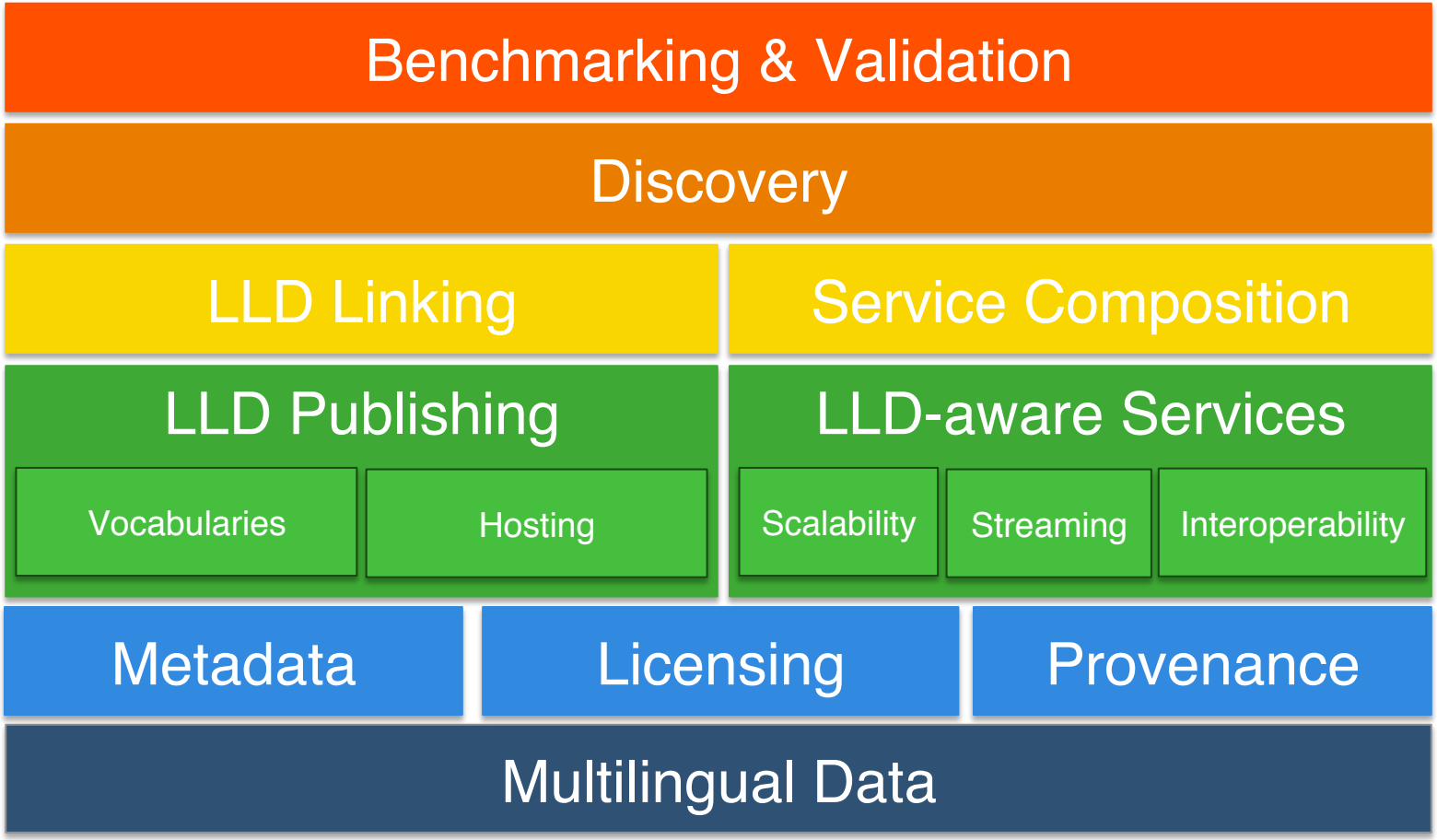
Reference Architecture

Discovery layer implemented by a number of independent indexing and aggregation services that support querying (SPARQL) and browsing data (Linked Data)

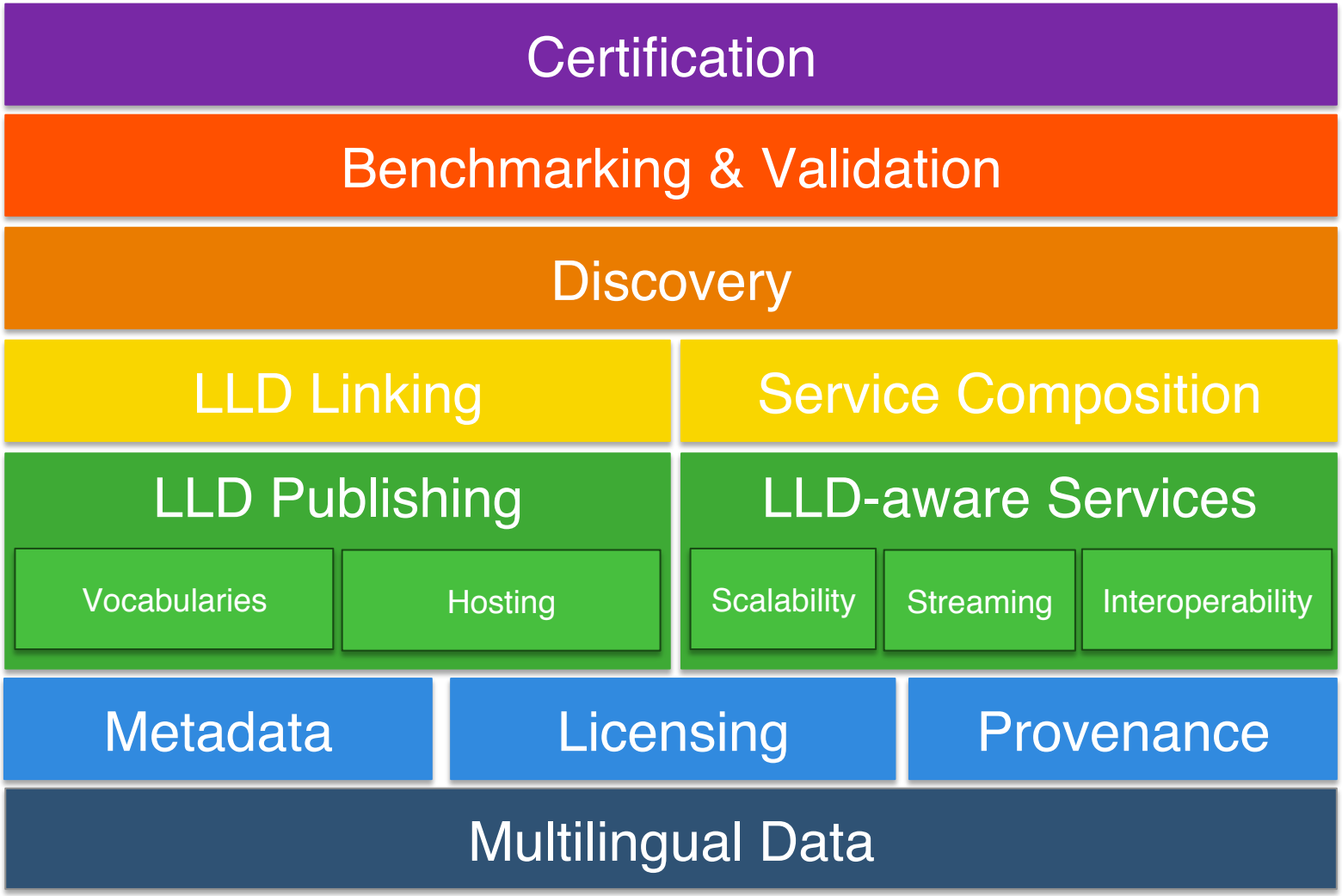


Reference Architecture

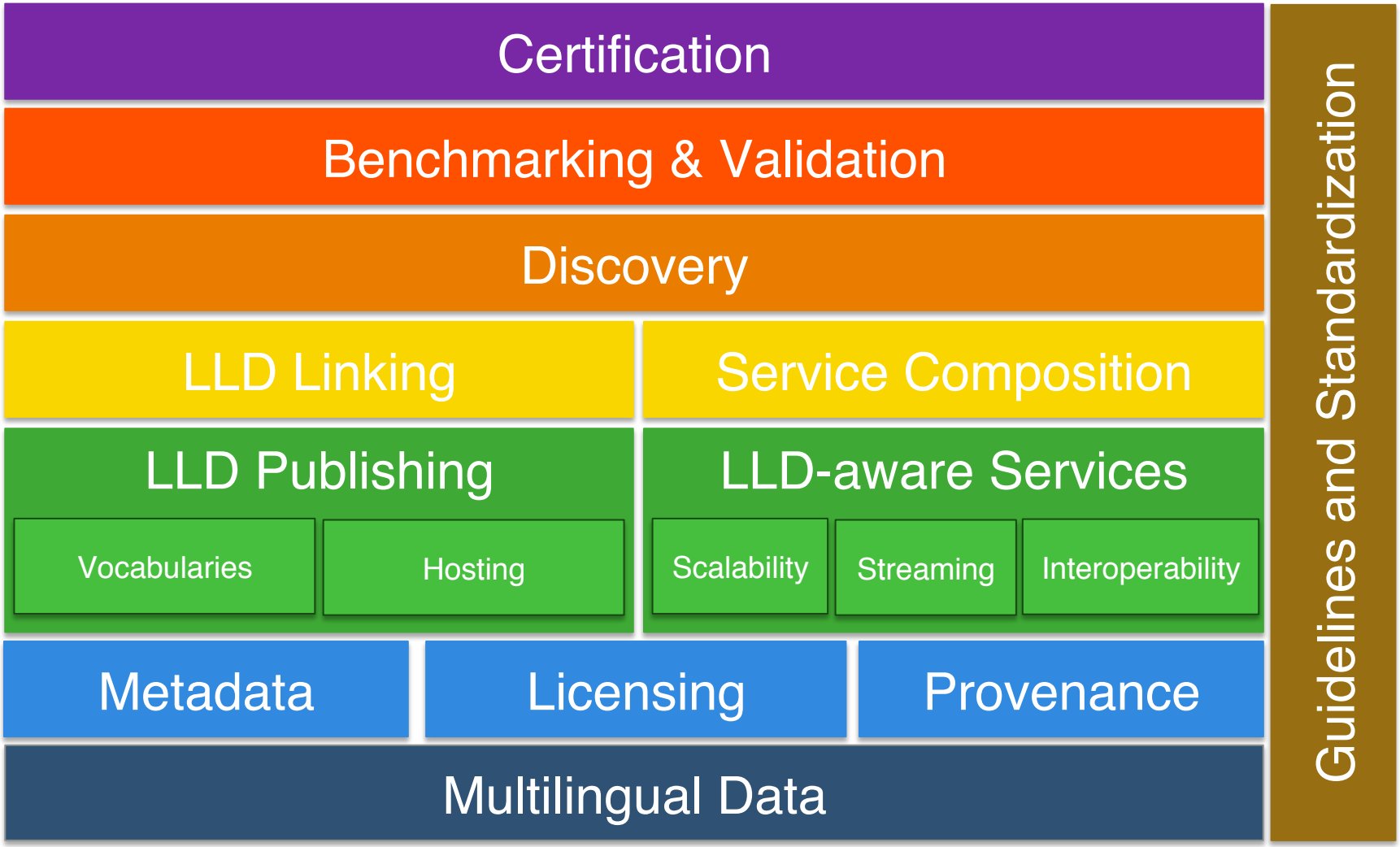
tools supporting comparison of datasets and services



Reference Architecture



Reference Architecture



Metadata

- **Metadata:** DataID for the description of datasets (see Reference Card for DataID), as well as Dublin Core, DCAT and a METASHARE ontology currently in development (see other threads)
- **Licensing:** The recommendation of the LIDER project is to use ODRL for the description of terms and conditions
- **Provenance:** The recommendation of the LIDER project is to use the PROV-O vocabulary to describe provenance of linguistic data resources Data Publishing: The LIDER project recommends to use DataHub for publishing metadata
- **Data Linking:** The LIDER project has implemented services that link data across sources as proof-of-concept implementation.

Discovery Layer

- Reference implementation is LingHub: http://linghub.lider-project.eu/
- Indexes metadata from METASHARE CLARIN, LRE Map, DataHub
- Integration and harmonization of data by mapping to DCAT, Dublin Core
- Exposes DataID metadata descriptions
- Provides SPARQL endpoint
- Browsable by humans and machines (Linked Data)

Source	Records	Triples	Triples per Record
META-SHARE	2,442	464,572	190.2
CLARIN	144,570	3,381,736	23.4
Datahub.io	218	10,739	49.3
LRE-Map	5,712	79,576	13.9

Property	Record Count (As percentage of all records)	Triples
Access URL	91,615 (91.6%)	191,006
Language	50,781 (50.7%)	98,267
Type	15,241 (15.2%)	17,894
Rights	3,080 (3.0%)	8915
Usage	3,397 (3.4%)	4,530

Services

Reference implementation of NLP services that:

- Use web sockets to process data in a streaming fashion
- Use NIF-grounded RDF/JSON-LD as input and output
- Can be composed together by merging output (RDF merge)

Standardization

Involvement in Community Groups:

- **Ontolex** (Ontology-Lexicon Models, CG)
- **BPMLOD** (Best Practices for Multilingual Linked Open Data, CG)
- **LD4LT** (Linked Data and Language Technologies, CG)

Use Cases

- An IT company is active in the **brand reputation market** and offers a product that is based on sentiment analysis for three languages (English, Spanish; Portuguese), and needs to find sentiment annotated data for German
- A **terminology management company** wants to exploit LLD to support the process of creating a corporate terminology. They want to provide seed terms and exploit LLOD to get further candidates for terms.
- A **machine translation company** wants to exploit LLOD for training machine translation system and ease the adaptation to a new domain, searches for parallel data on a certain language pair.
- An IT company develops information extraction techniques for **competitor analysis**. It needs to develop an application that works on Twitter data. The company needs to find POS-annotated Twitter data to adapt their POS tagger to the Twitter domain.
- A **researcher** wants to publish a dataset on the Web as Linguistic Linked Data and needs support in this. A part of the dataset will be offered for free and part will be offered in exchange of money.

Discussion

Thanks for your attention!
Any comments, questions,
…?