OpenData.sk

Výzvy a technické riešenia

Rastislav Senderák, EEA

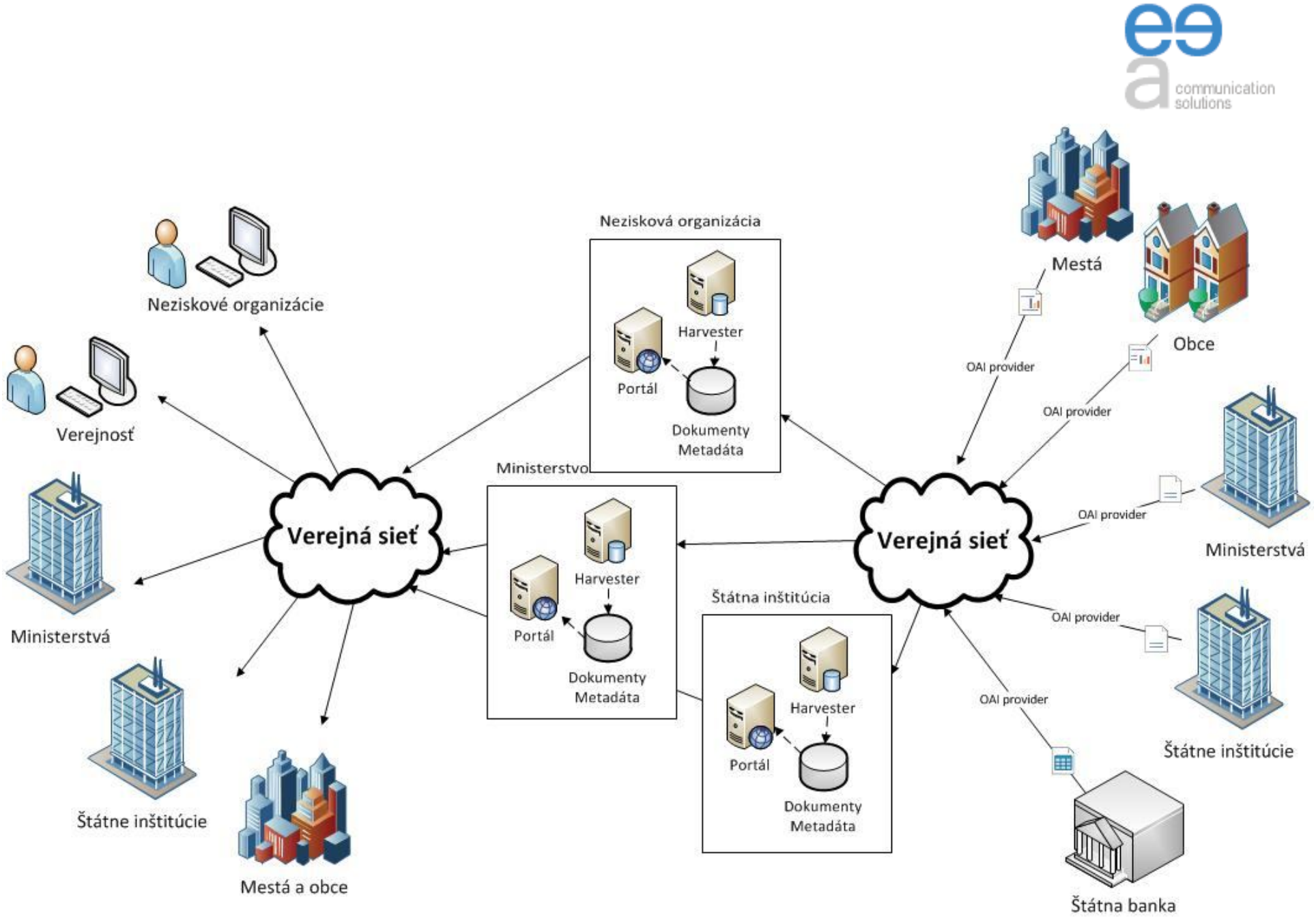
ITAPA2011, 25.11.2011

Agenda

- Ciele
- Výzvy
- Prístup k riešeniu
- Architektúra

Ciele

- Technologická podpora iniciatívy OpenData
  - Zabezpečiť potrebnú IT a metodickú infraštruktúru
  - Výber komponentov riešenia
  - Zostavenie OpenSource balíka a metodiky jeho nasadenia


Neziskové organizácie
Verejnosť
Ministerstvá
Štátne inštitúcie
Mestá a obce
Verejná sieť
Nezisková organizácia
Harvester
Portál
Dokumenty
Metadáta
Ministerstvo
Harvester
Portál
Dokumenty
Metadáta
Štátna inštitúcia
Harvester
Portál
Dokumenty
Metadáta
Verejná sieť
Mestá
Obce
OAI provider
OAI provider
OAI provider
Ministerstvá
OAI provider
Štátne inštitúcie
OAI provider
Štátna banka

Výzvy

- ako zbierať
  - push, pull?
- ako spracovať, ako ukladať
- ako poskytovať
  - princípy, služby, formáty, protokoly
- ako organizovať
  - centrálne vs. distribuovane,
  - zdola nahor vs. zhora nadol, ...

6

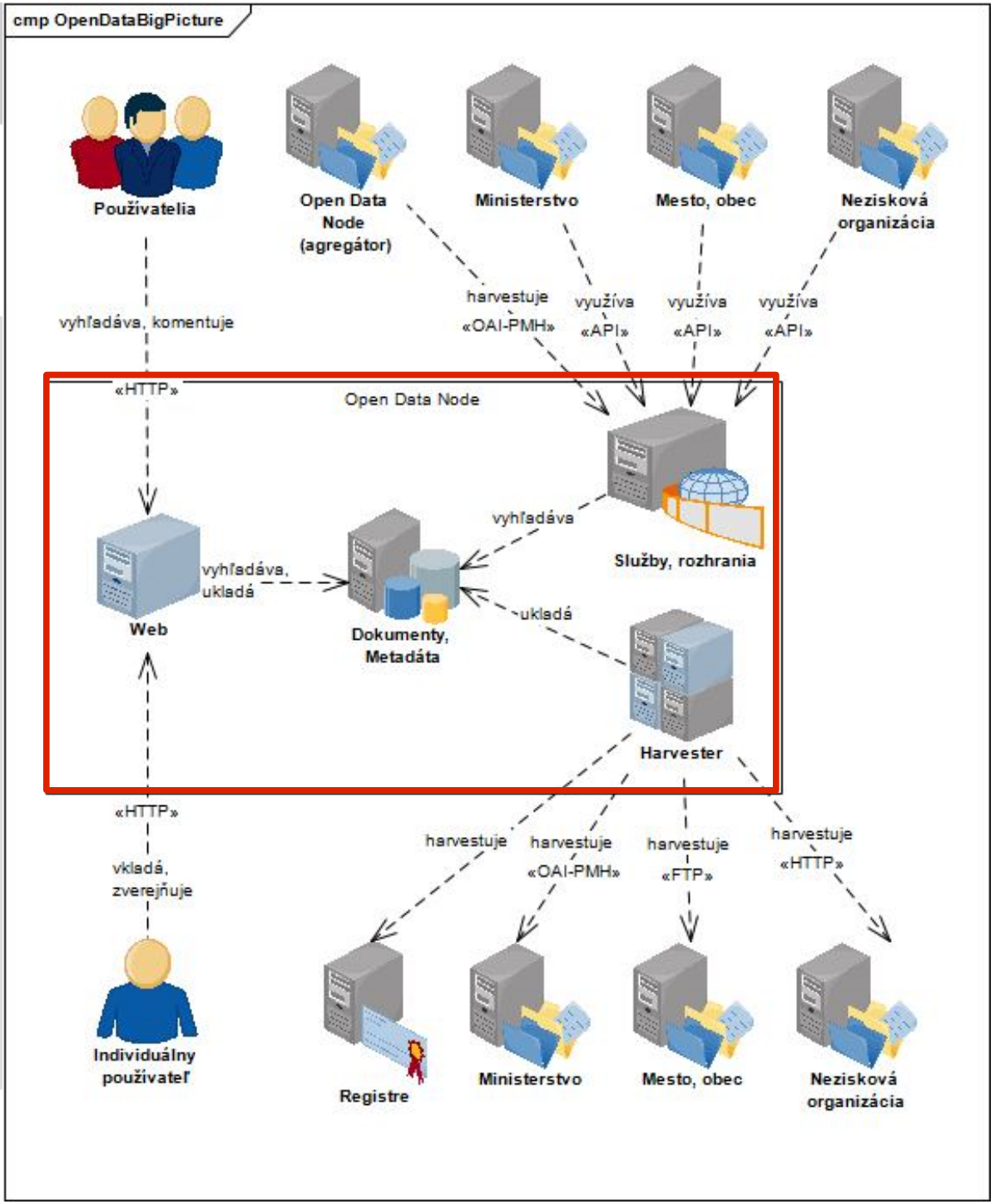
Open Data Node

Nainštalovaný balík

- zbiera dokumenty

a metadáta,

- spracúva ich
- a poskytuje
  - verejnosti,
  - inštitúciám,
  - aplikáciám...

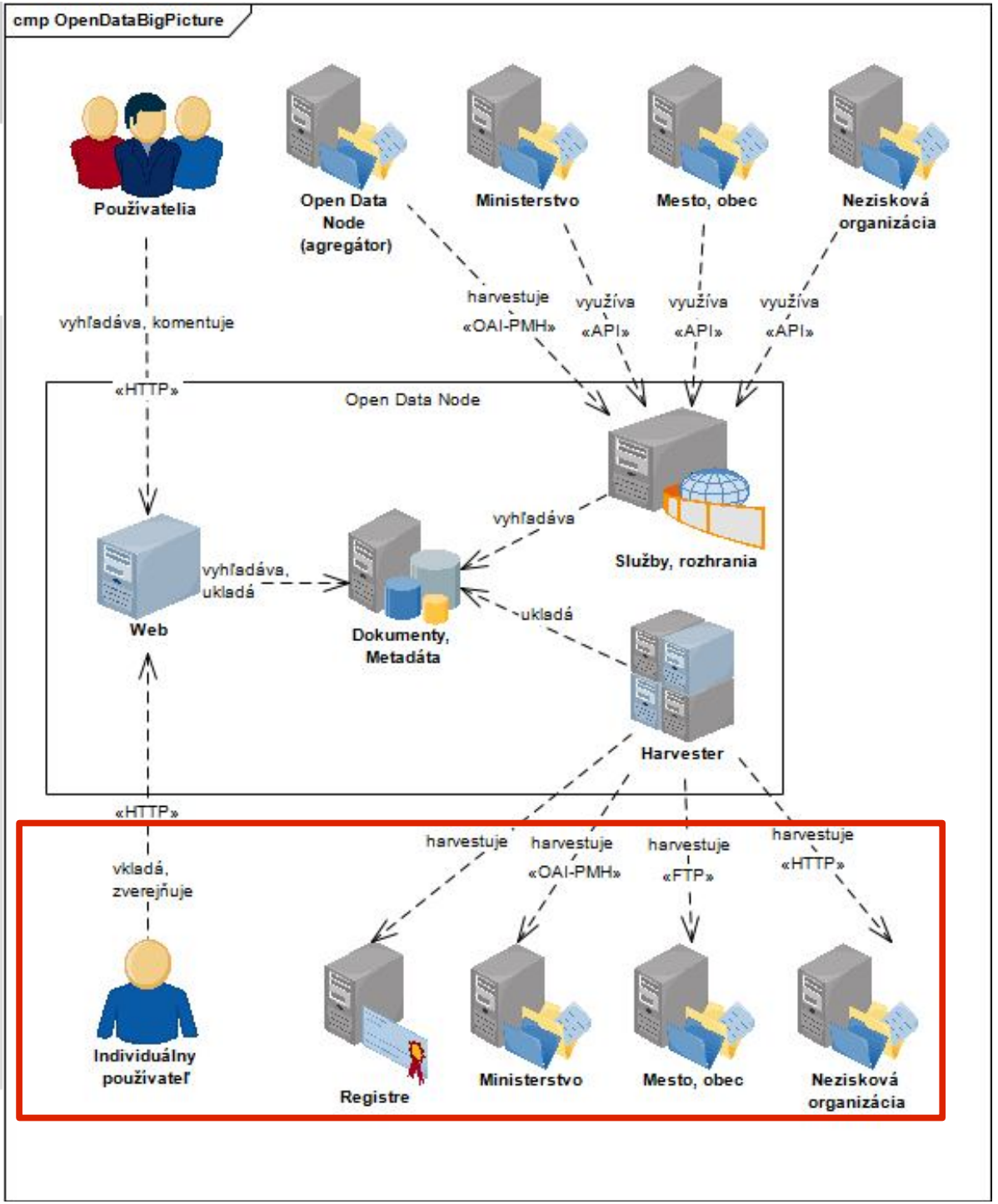


7

Dáta

Ako zbierať?

- harvesting
- manuálne / form
- enrichment
- anotácie

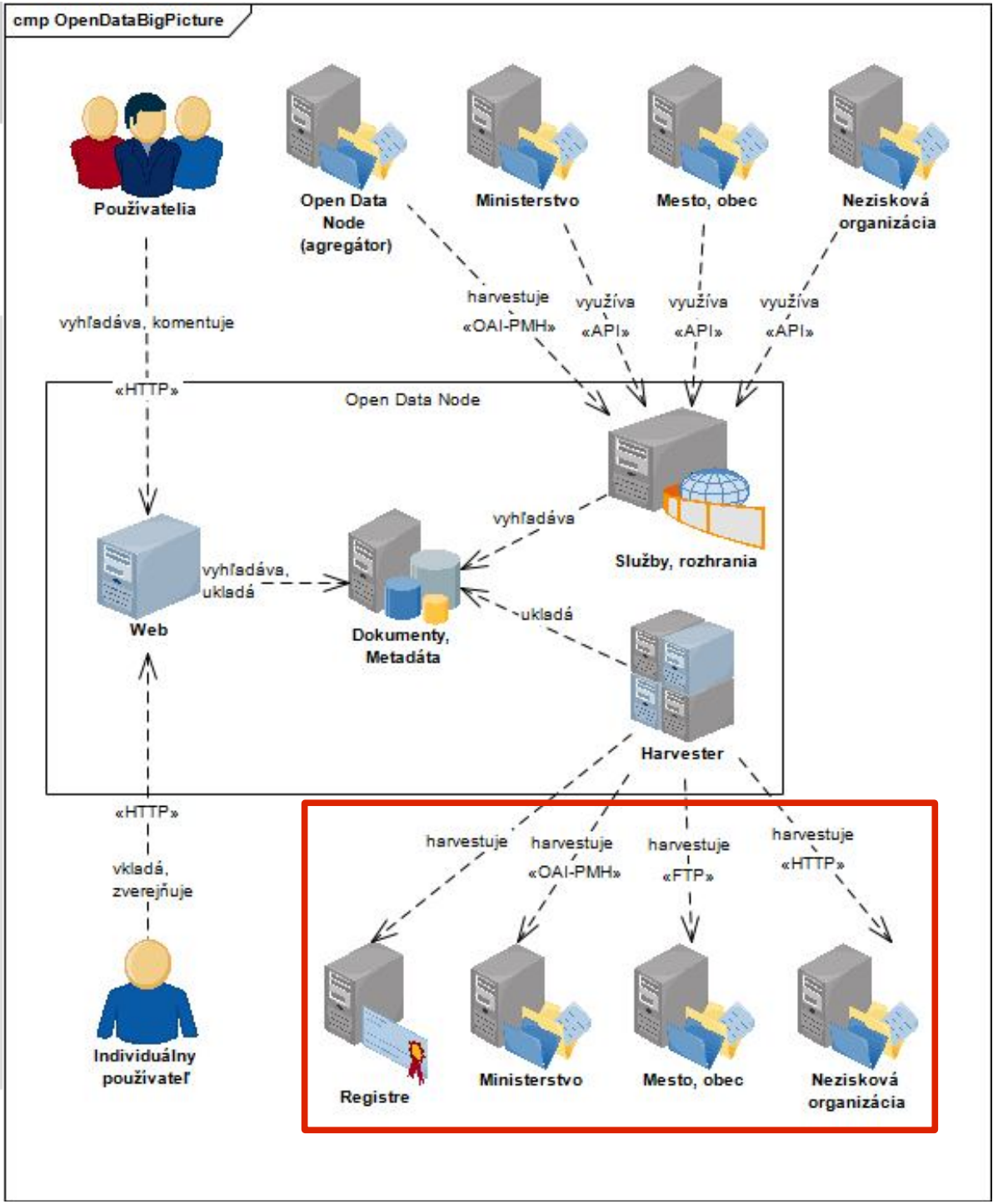


8

Dáta

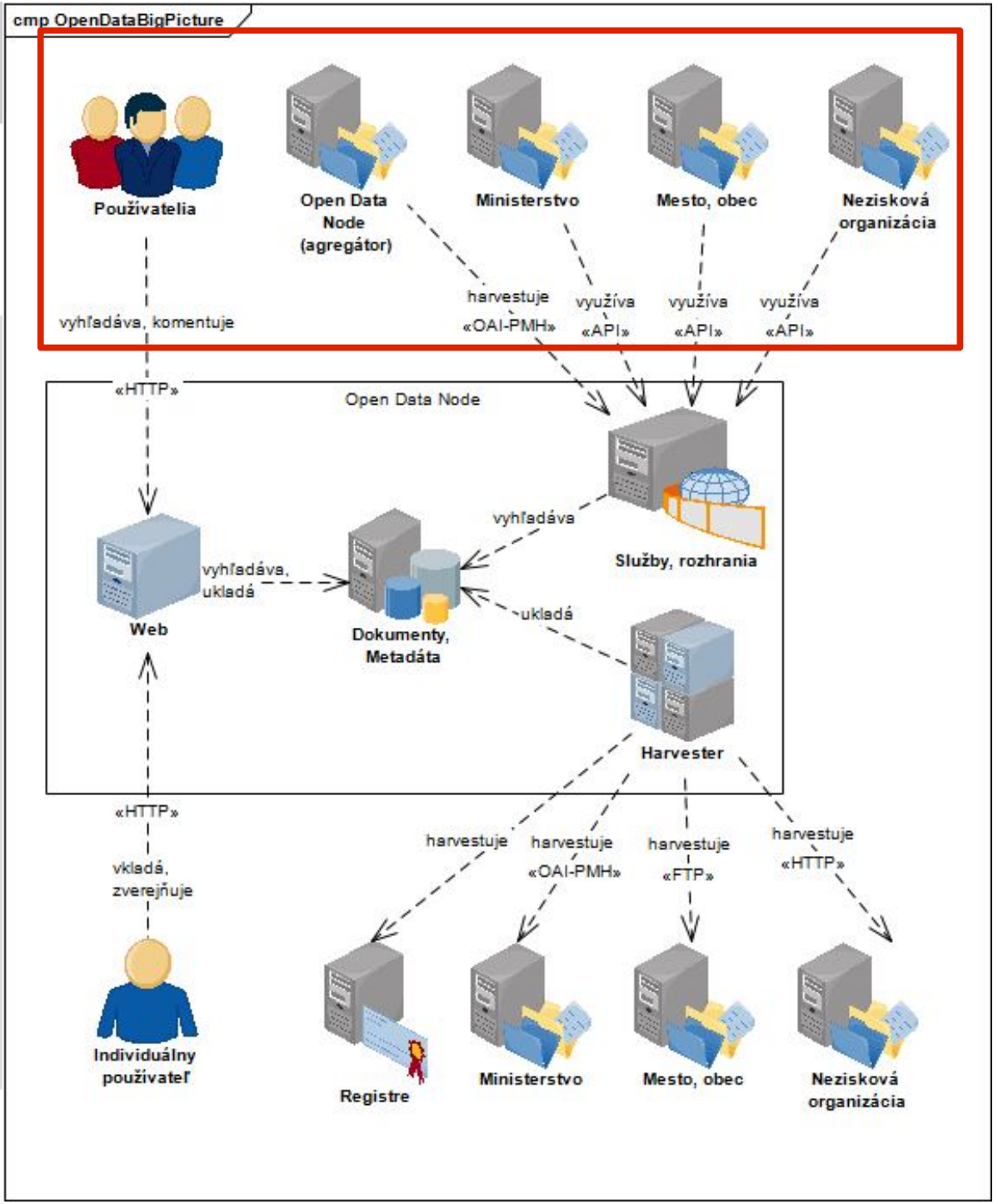
Aké zdroje?

- IS inštitúcií – verejná správa, štátna správa, tretí sektor, súkromné zdroje
- registre



Sprístupnenie

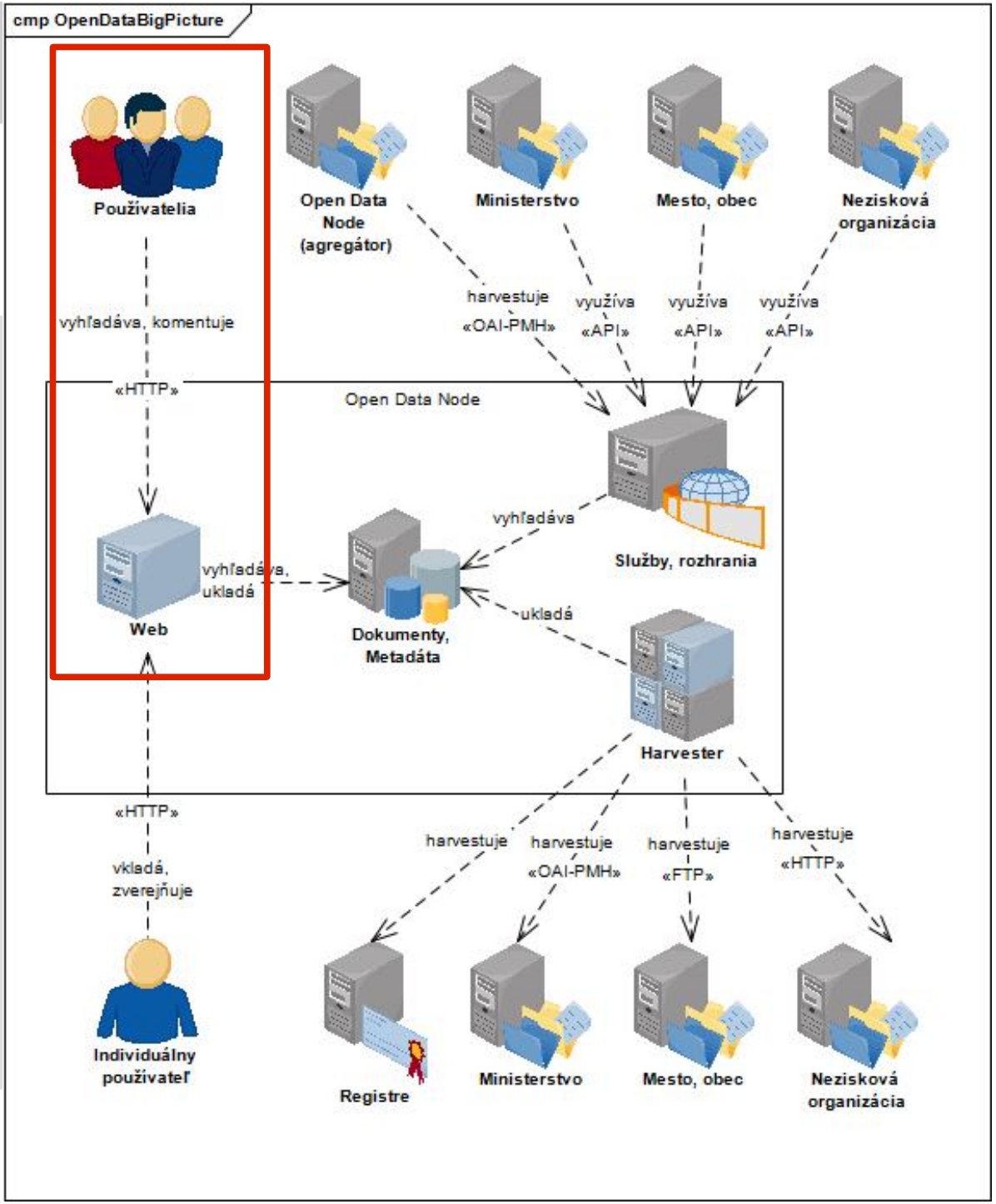
- pre harvesting
- pre aplikácie
  - SOAP, REST...
  - XML, JSON...
- ako služby, portál, aplikácie
- ako dátový export



10

Prezentácia

- zoznam
- fazety
- mapa
- časová os
- tabuľky
- graf

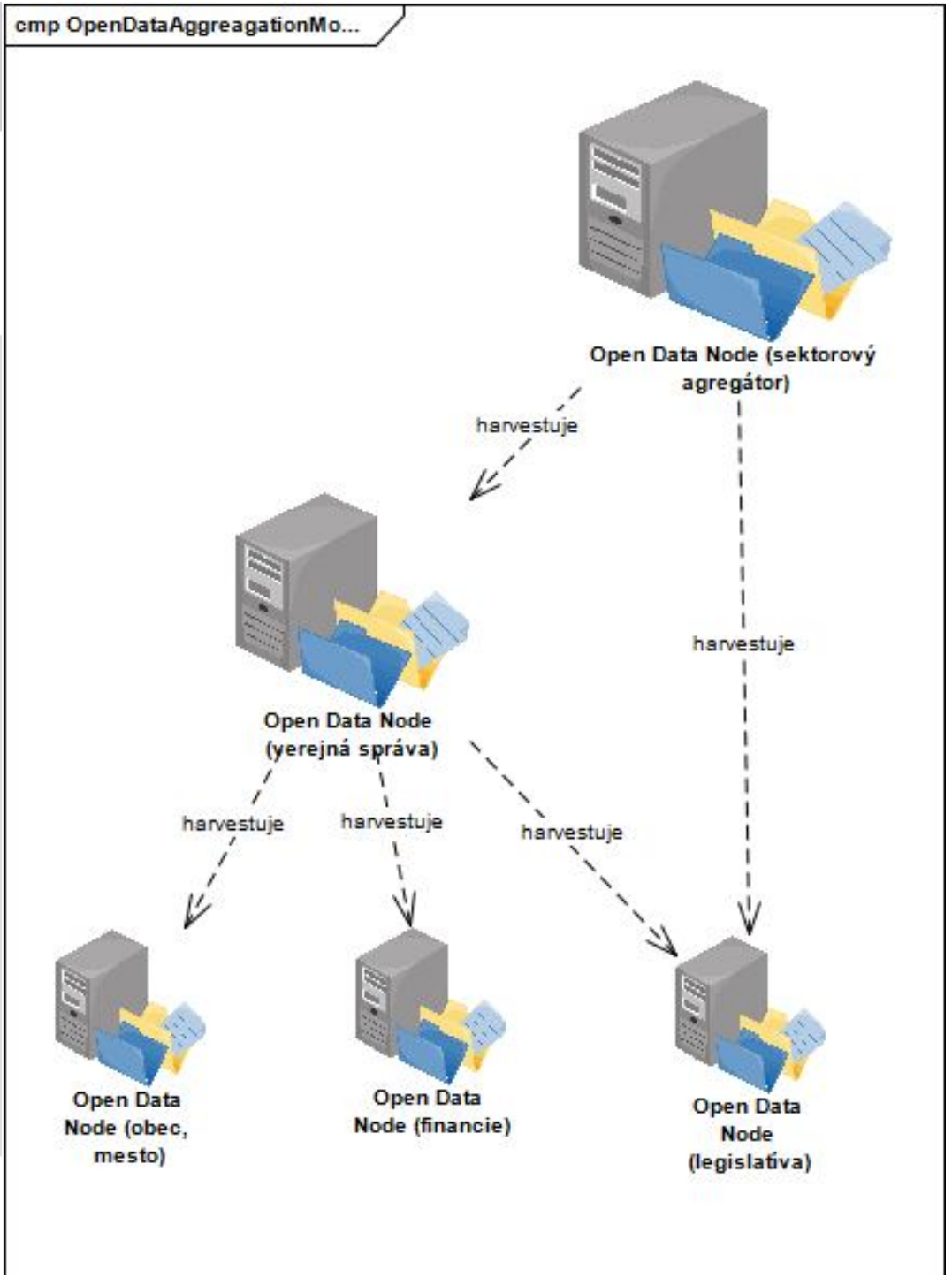


Otvorená architektúra

- autonómne,
- kaskády / hierarchie

Podľa

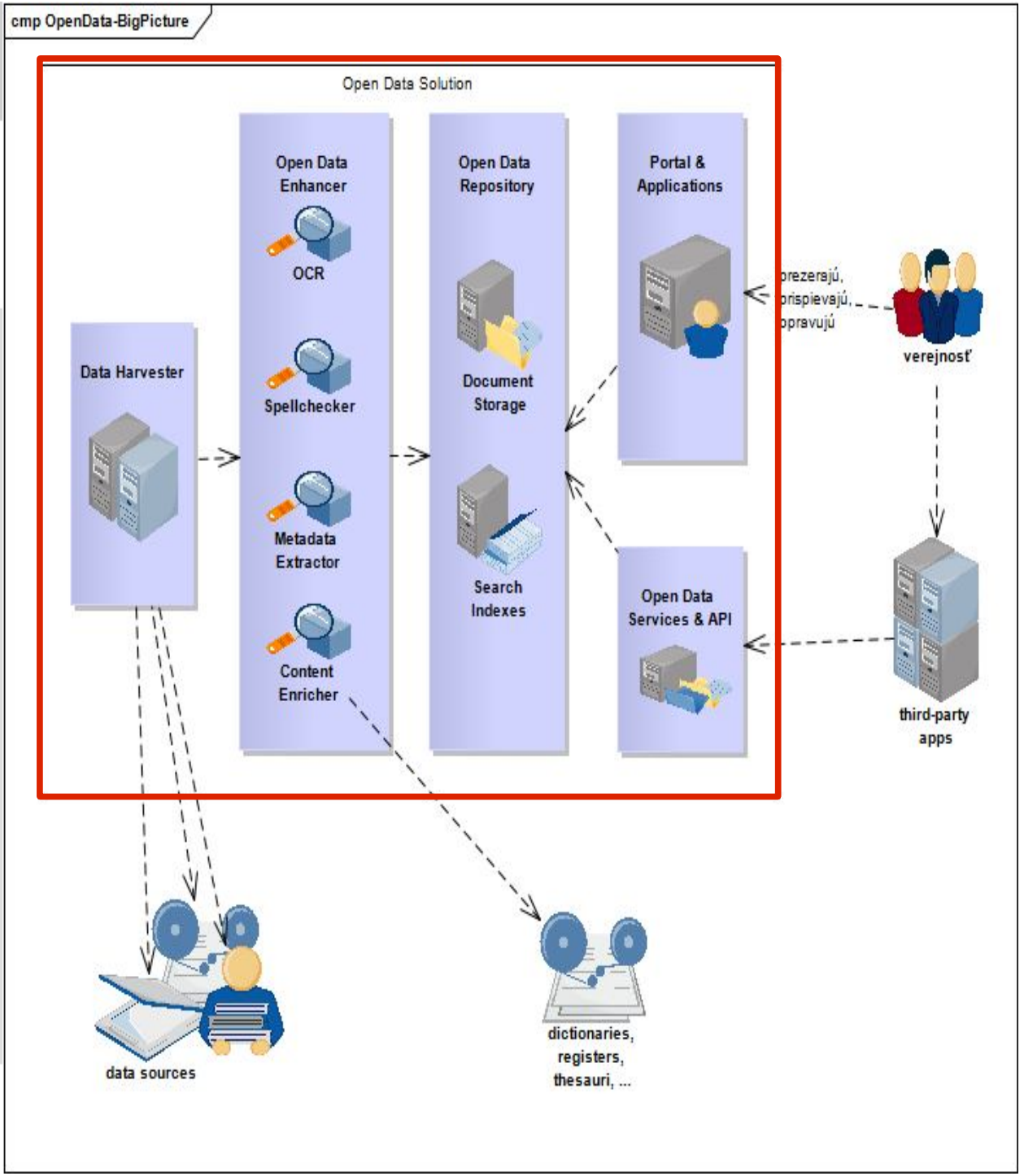
- domény,
- komunity,
- geograficky,...



12

Open Data Node

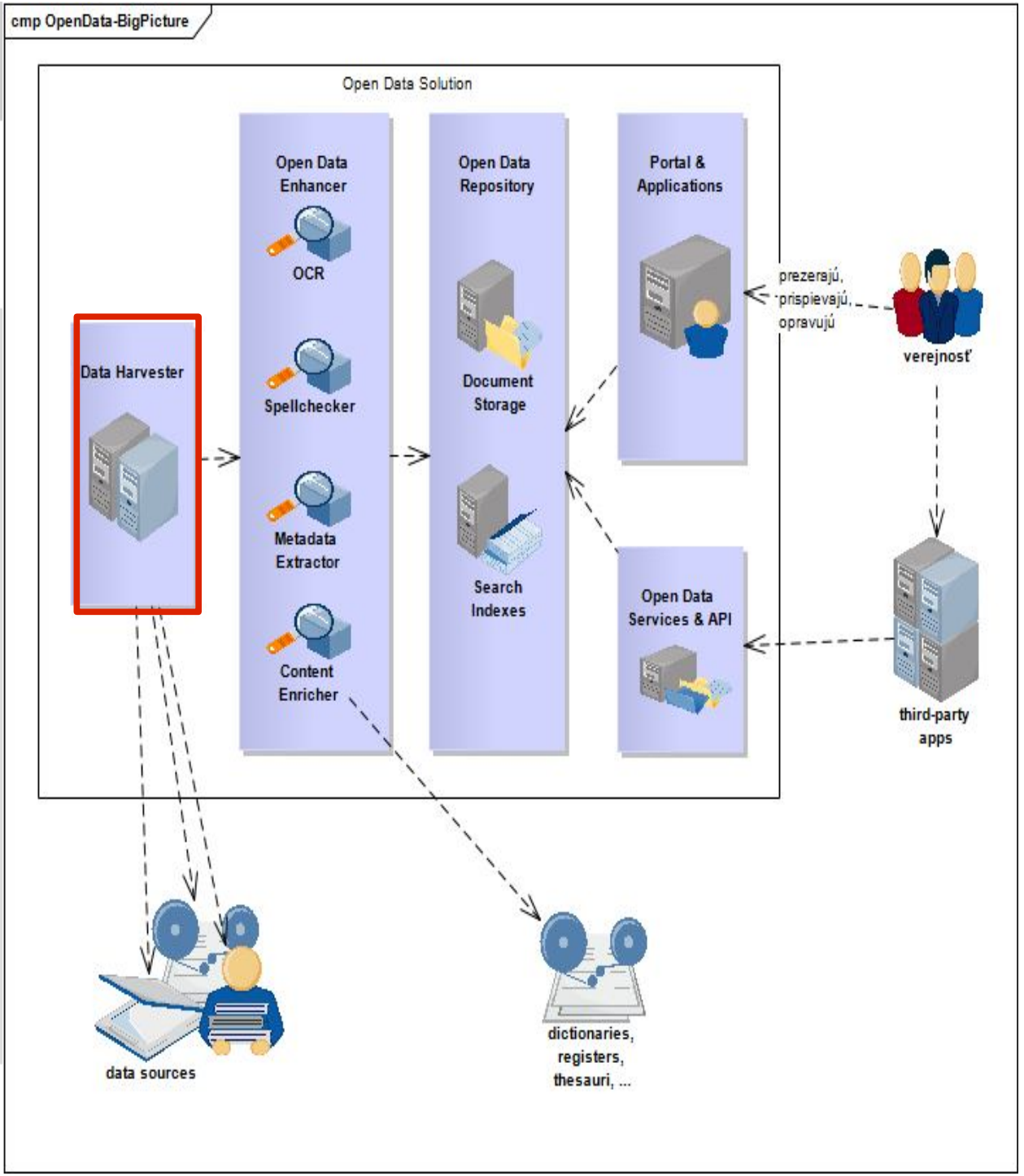
- Harvester
- Enhancer
- Repo
- Services/API
- Portál
- Aplikácie



13

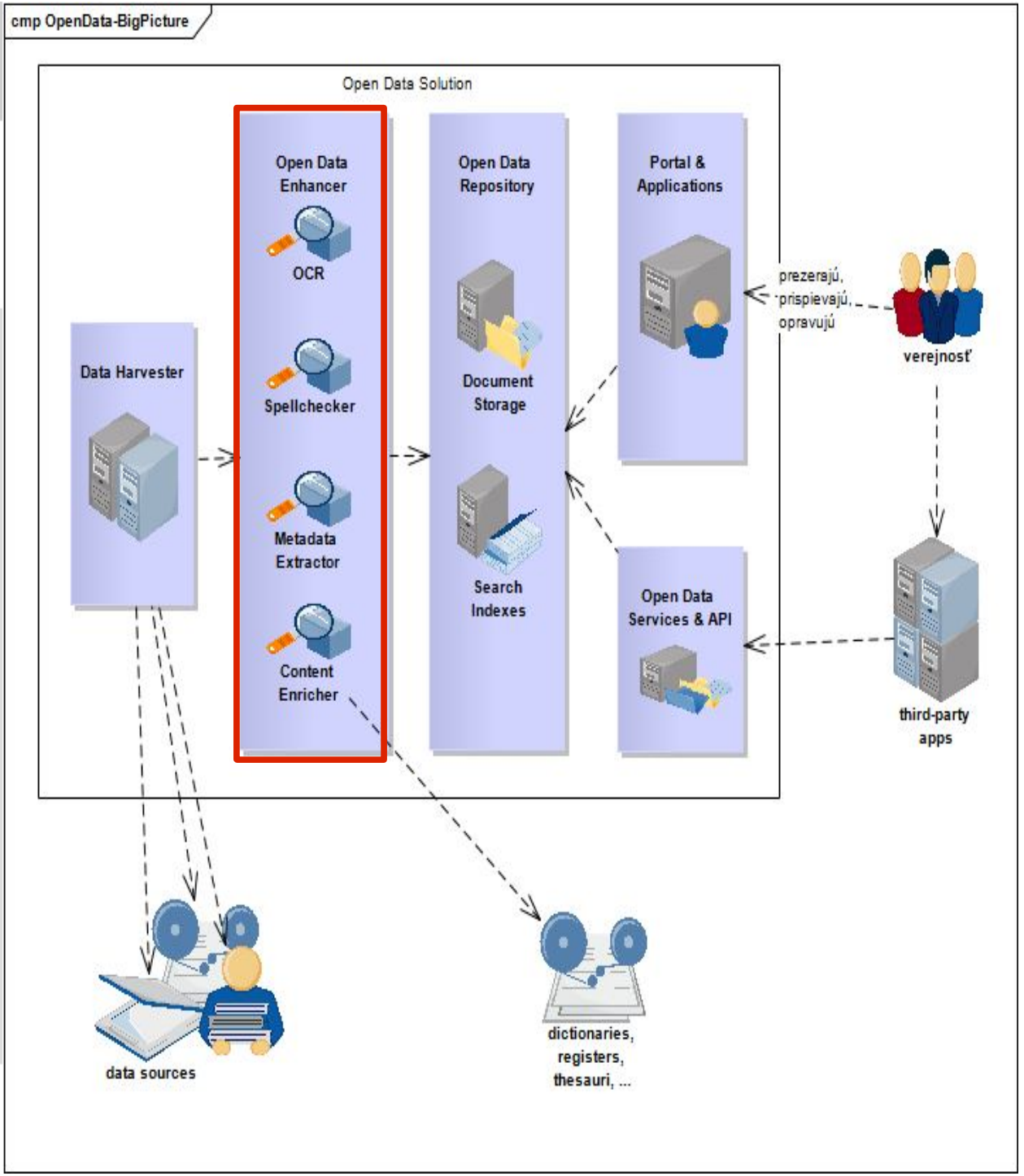
Harvester

- zber dát
  - OAI-PMH, HTTP, FTP, ...
- validácia, transformácia
- technické metadáta
- scheduler



Enhancer

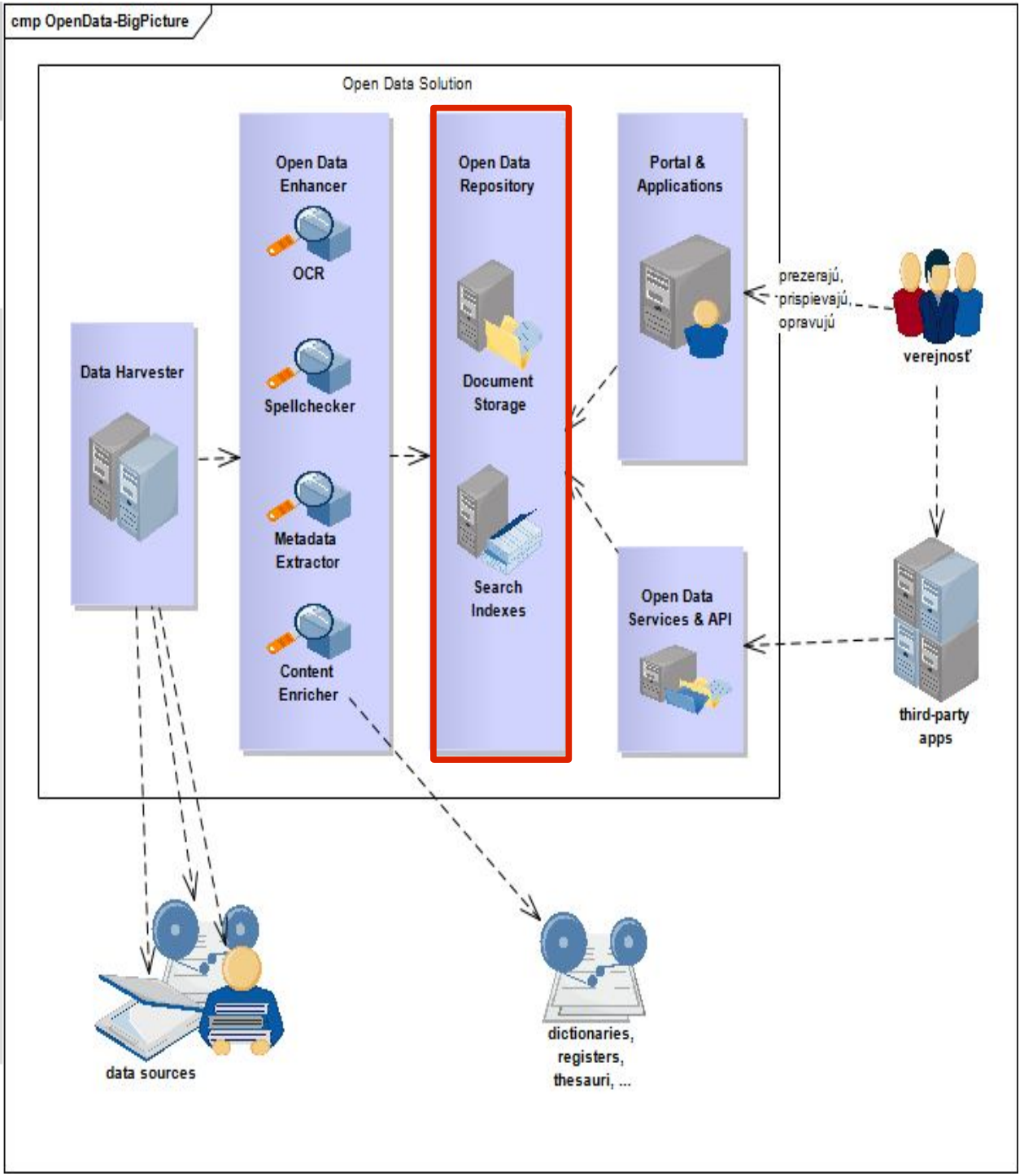
- metadáta!
  - OCR
  - text mining
- kvalita, úplnosť,...



15

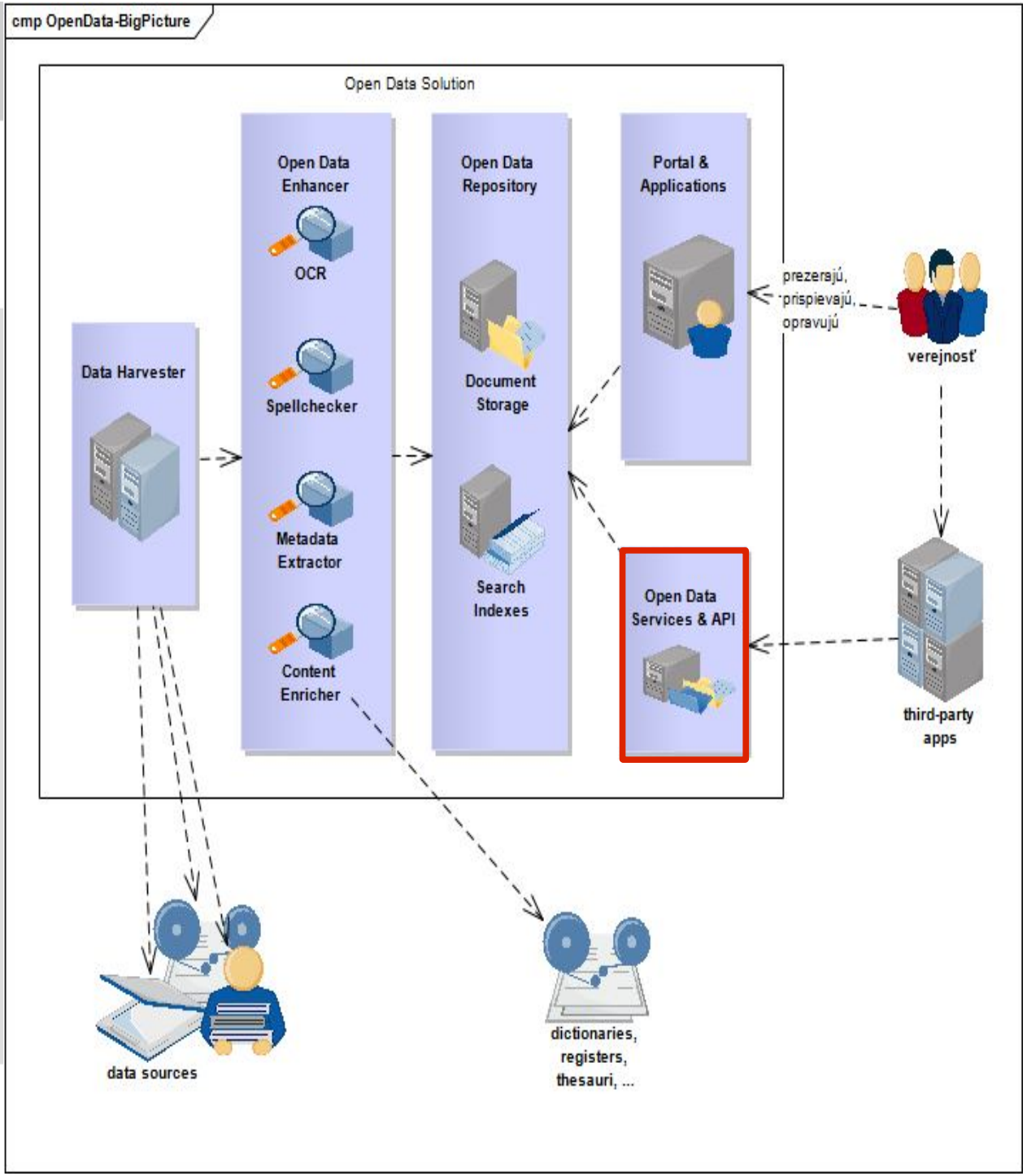
Repo

- úložisko
  - dáta
  - metadáta
  - indexy...
  - IS / SS
- identifikácia, deduplikácia,...
- monitoring



Služby

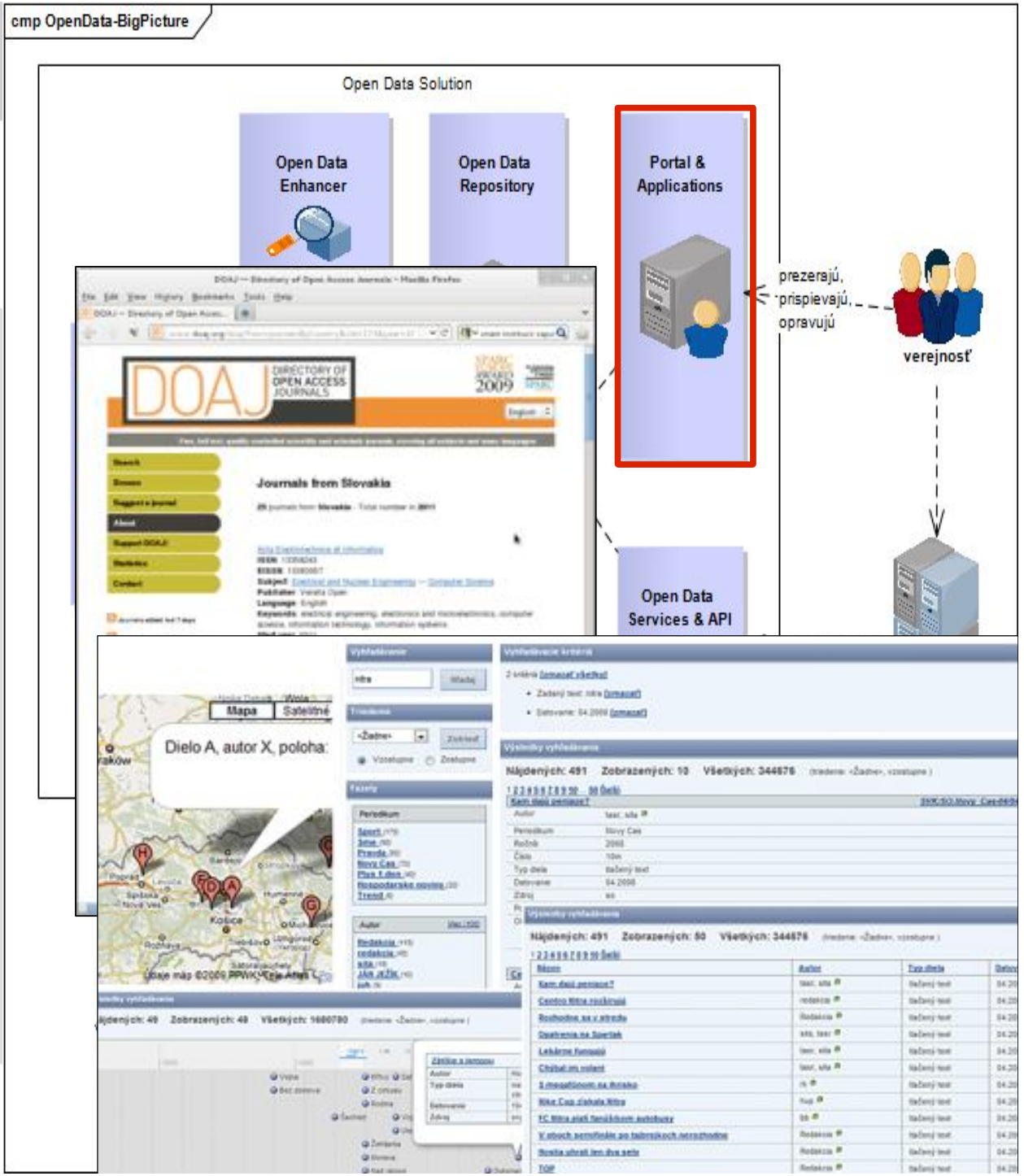
- API
  - protokoly
  - formáty
  - exporty



17

Portál

- info
- dokumenty
  - vyhľadáva
  - prezentuje
- komunita, kolaborácia, crowdsourcing
- dokumentácia
- administrácia



Komponenty

- OAI-PMH Harvester
- Enhancer
  - Tesseract OCR / Ephesoft Document Capture
- Repo
  - Apache Jackrabbit (JCR)
  - Apache SOLR (indexy)
  - Djatoka Image Server / IIPImage
  - PostgreSQL
- Portál: Liferay

Parametre

- Repo: 300 mil. záznamov (cca 4 miliardy tripletov)
- Repo: > 2 TB digitálneho obsahu
- Harvester: 30 (nových) záznamov za sekundu
- Portál: vyhľadanie do 0.5 sekúnd (pre 1 CU) resp. do 2 sekúnd (pre 50 CU)

ea communication solutions

Ďakujem