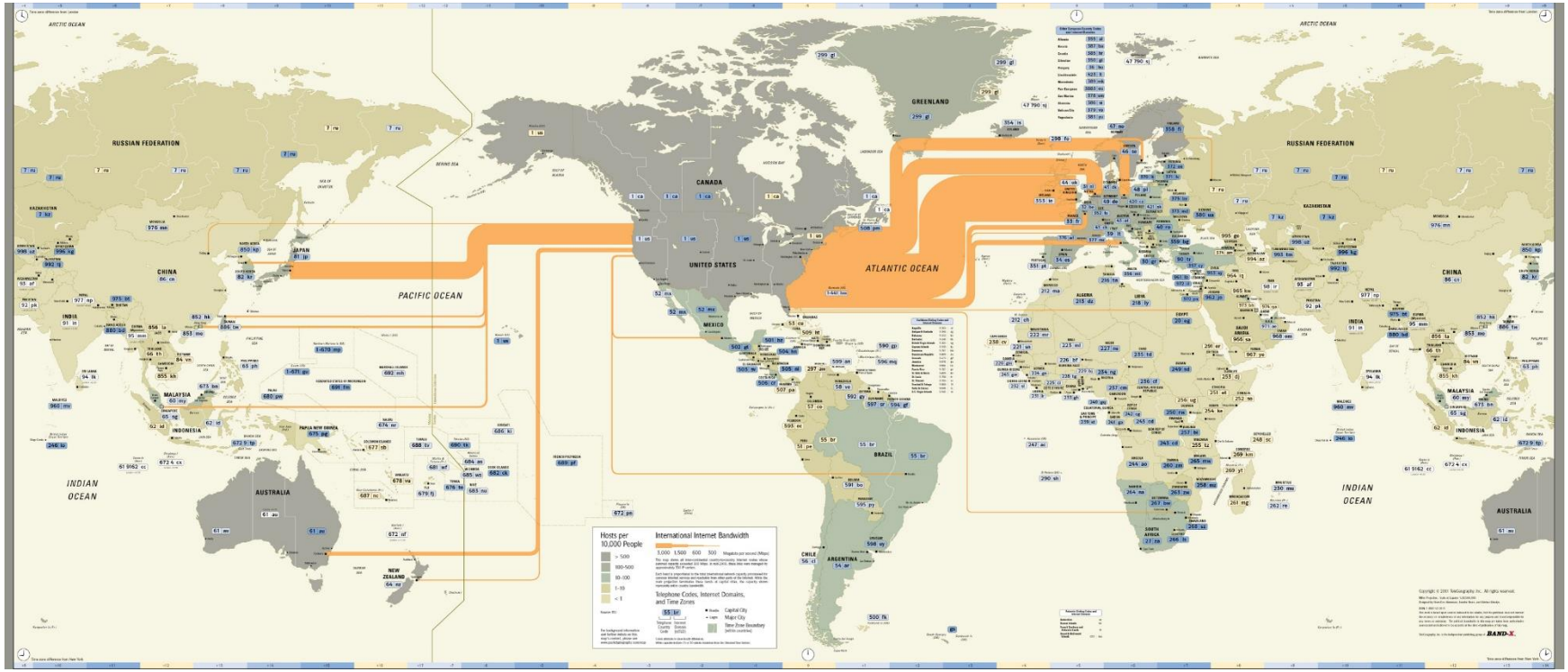
HackGeoLing

Hackathon Geolinguistic Crossreference Miner

ODER: Eine Homage an UTF-8 und PHP

Joseph Birkner, Tillmann Dönicke | #clunc15

Idee

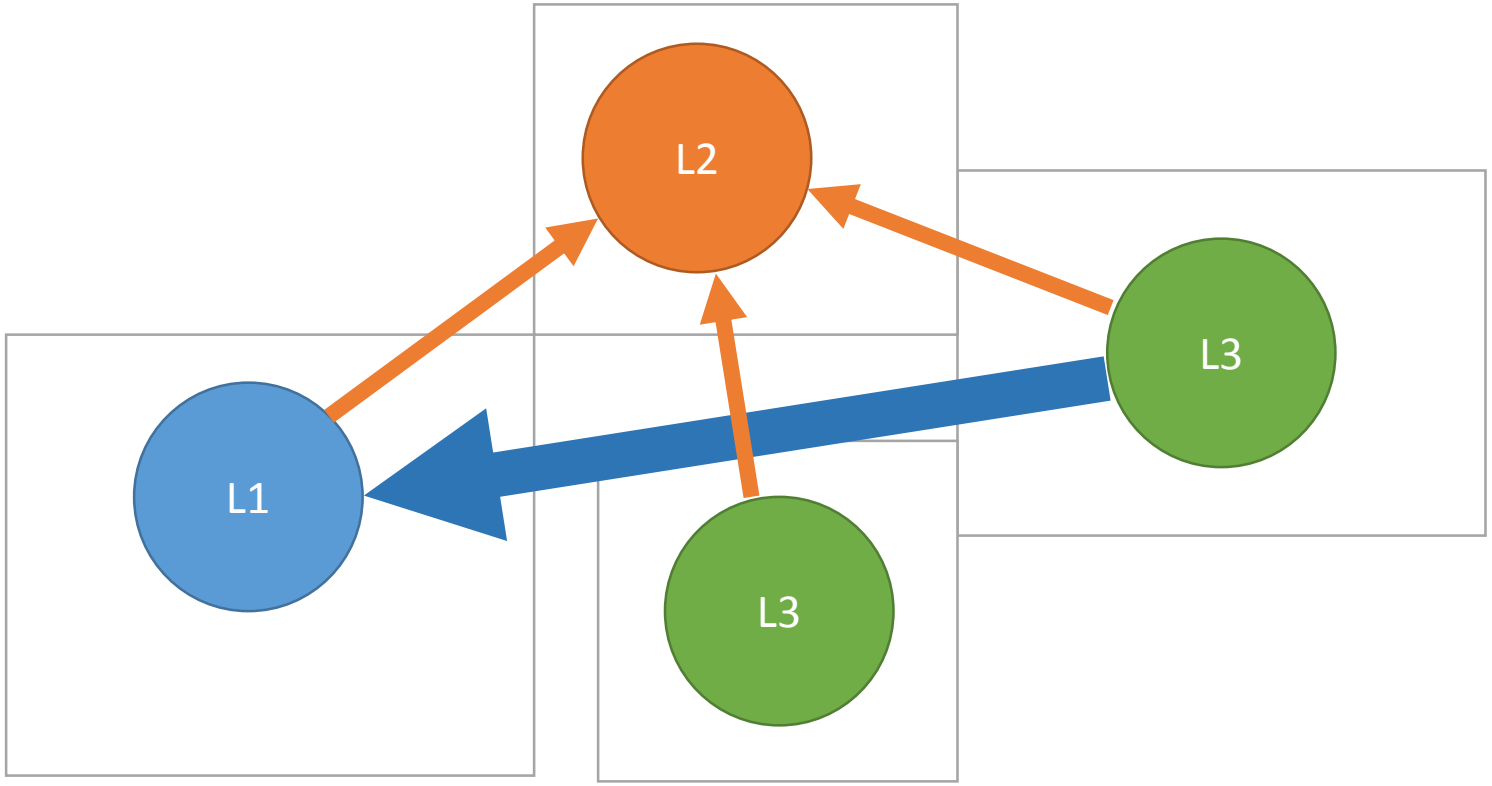

ARCTIC OCEAN
RUSSIAN FEDERATION
GREENLAND
CANADA
UNITED STATES
ATLANTIC OCEAN
PACIFIC OCEAN
MEXICO
CHINA
JAPAN
INDIA
MALAYSIA
INDONESIA
AUSTRALIA
NEW ZEALAND
BRAZIL
ARGENTINA
CHILE
INDIAN OCEAN
SOUTH AFRICA
Hosts per 10,000 People
International Internet Bandwidth
Telephone Codes, Internet Domains, and Time Zones

Idee

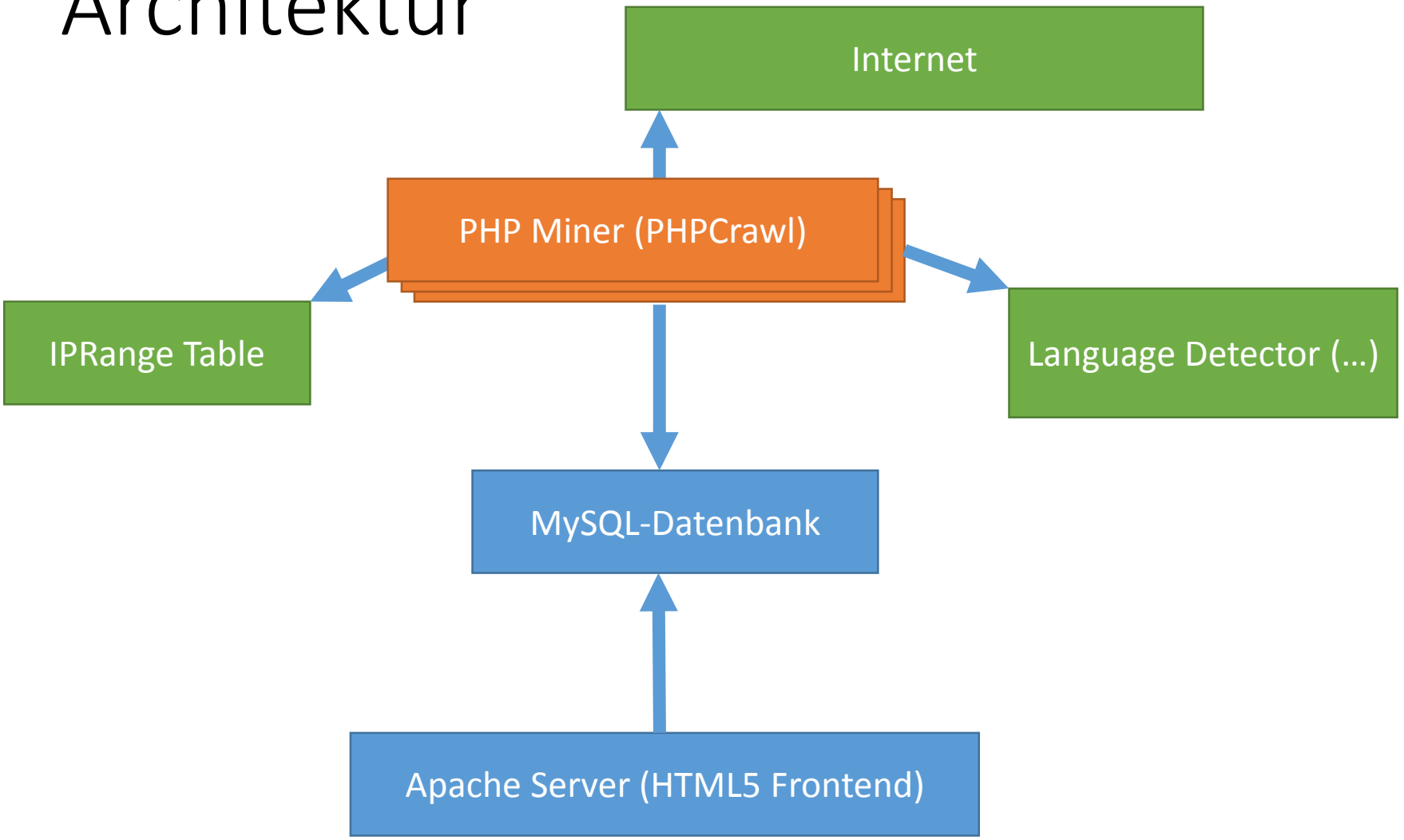
- „Linguistische Karte“ des Internets
- Welche Sprachen verlinken in welchem Maße aufeinander?
- Welche Sprachen verlinken vornehmlich auf sich selbst/untereinander?
- Wie unterscheiden sich diese Verhältnisse in verschiedenen Kontexten (Suchbegriffe)?

Daten

Startsprache	Zielsprache	Startland	Zielland	Suchbegriff



Architektur



Plans are made to be….

- Language Detector:
  - PHP TextCat v0.5 (Alles chinesisch)
  - Java Apache Tika (Alles <lt>?)
  - WTF-8?
- IP-Adressen
  - PHP ist auf Windows auf 32b Signed Int beschränkt

Crawler-Konfiguration

- Nur <a href…> folgen
- Auf 15 Follows pro Host beschränkt
- (Performance…) (Bots running on localhost)

Demo!

- …

Request for Comment

- How to deal with encodings on the WWW in relation to language detection?
- Which language detector to use?
- Which crawler APIs to use?

- Request for help!!