Beacon Network: Systém pre globálne zdieľanie genomických dát

Miroslav Cupák

Diplomová práca

15. 2. 2016

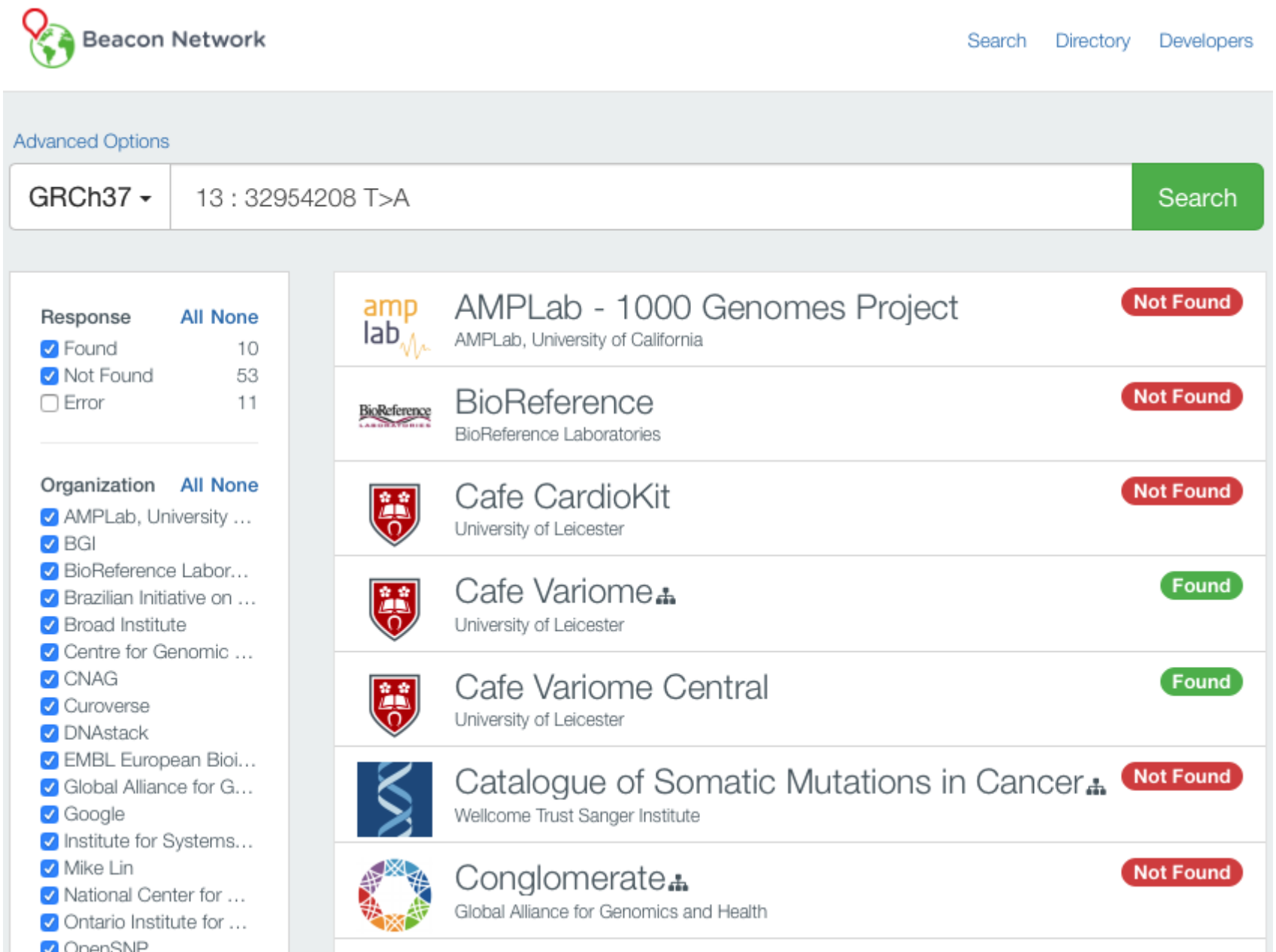
Problém

- cena sekvenovania genómu klesla za posledných niekoľko rokov miliónnásobne
- množstvo sekvenovaných dát sa zdvojnásobuje každých 7 mesiacov
- do roku 2025 bude sekvenovaných 10^{21} bázových párov (100 miliónov až 2 miliardy ľudí)
- dáta analyzované v izolácii podľa populácie, choroby, organizácie a pod., čo nestačí
- riešenie
  - zdieľanie dát medzi organizáciami (**Beacon Project**)
  - federatívny systém na agregáciu informácií z rôznych organizácií (**Beacon Network**)

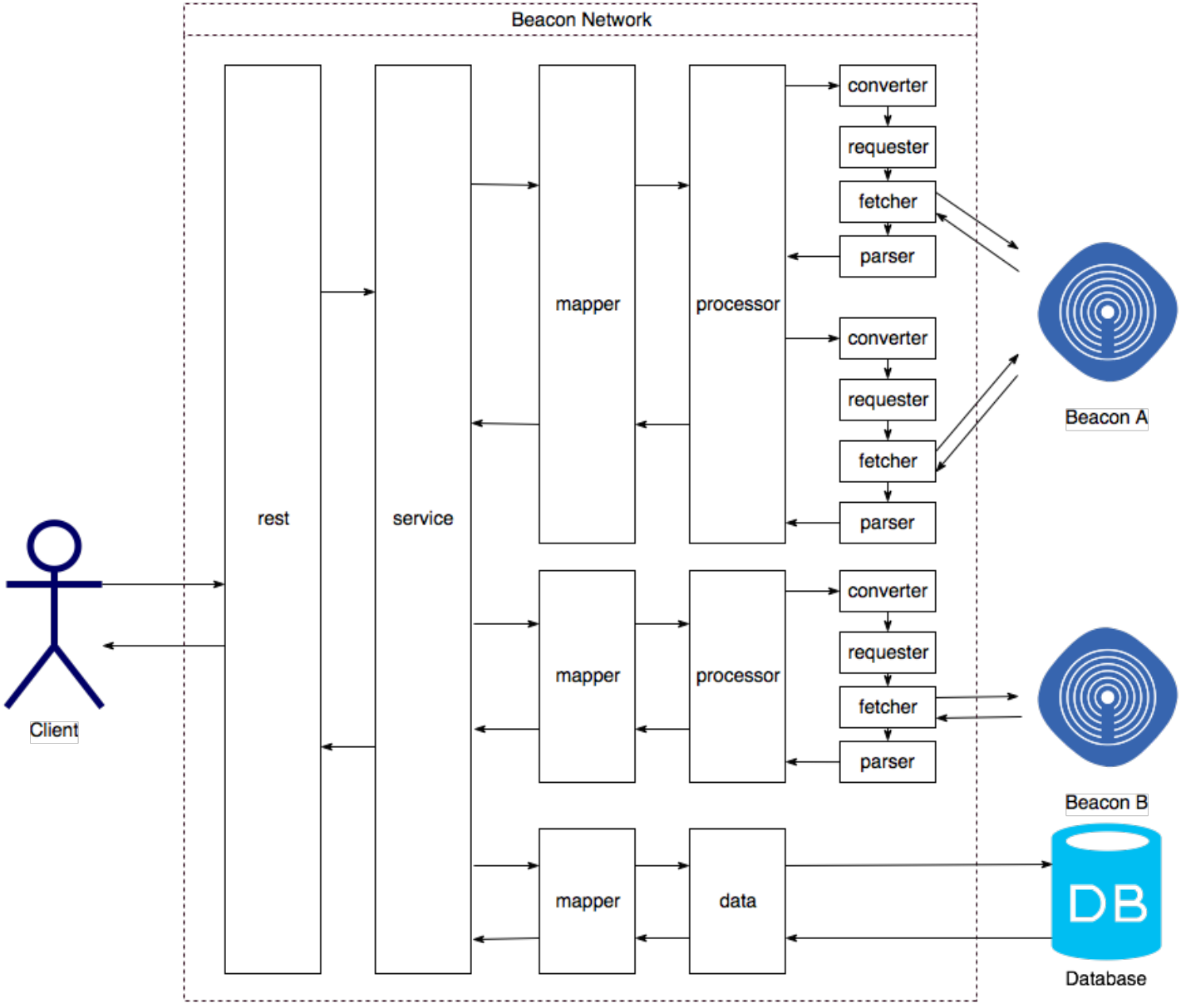
Beacon Project

- Globálna aliancia pre genetiku a zdravie (GA4GH)
- demonštračný projekt
- test ochoty inštitúcií zdieľať genetické informácie
- webová služba zisťujúca výskyt genetickej variácie v databáze inštitúcie
- otázka: *Existuje A na pozícii 100 000 v chromozóme 1?*
- odpoveď: *áno/nie*
- problémy v 2014:
  - neexistuje špecifikácia
  - 4 implementácie s rôznymi parametrami a ich hodnotami, metódami prístupu, formátmi...

Klient


Beacon Network
Search
Directory
Developers
Advanced Options
GRCh37
13 : 32954208 T>A
Search
Response All None
Found 10
Not Found 53
Error 11
Organization All None
AMPLab, University ...
BGI
BioReference Labor...
Brazilian Initiative on ...
Broad Institute
Centre for Genomic ...
CNAG
Curoverse
DNAstack
EMBL European Bioi...
Global Alliance for G...
Google
Institute for Systems...
Mike Lin
National Center for ...
Ontario Institute for ...
OpenSNP
AMPLab - 1000 Genomes Project
AMPLab, University of California
Not Found
BioReference
BioReference Laboratories
Not Found
Cafe CardioKit
University of Leicester
Not Found
Cafe Variome
University of Leicester
Found
Cafe Variome Central
University of Leicester
Found
Catalogue of Somatic Mutations in Cancer
Wellcome Trust Sanger Institute
Not Found
Conglomerate
Global Alliance for Genomics and Health
Not Found

Spracovanie dotazu



Podporné nástroje

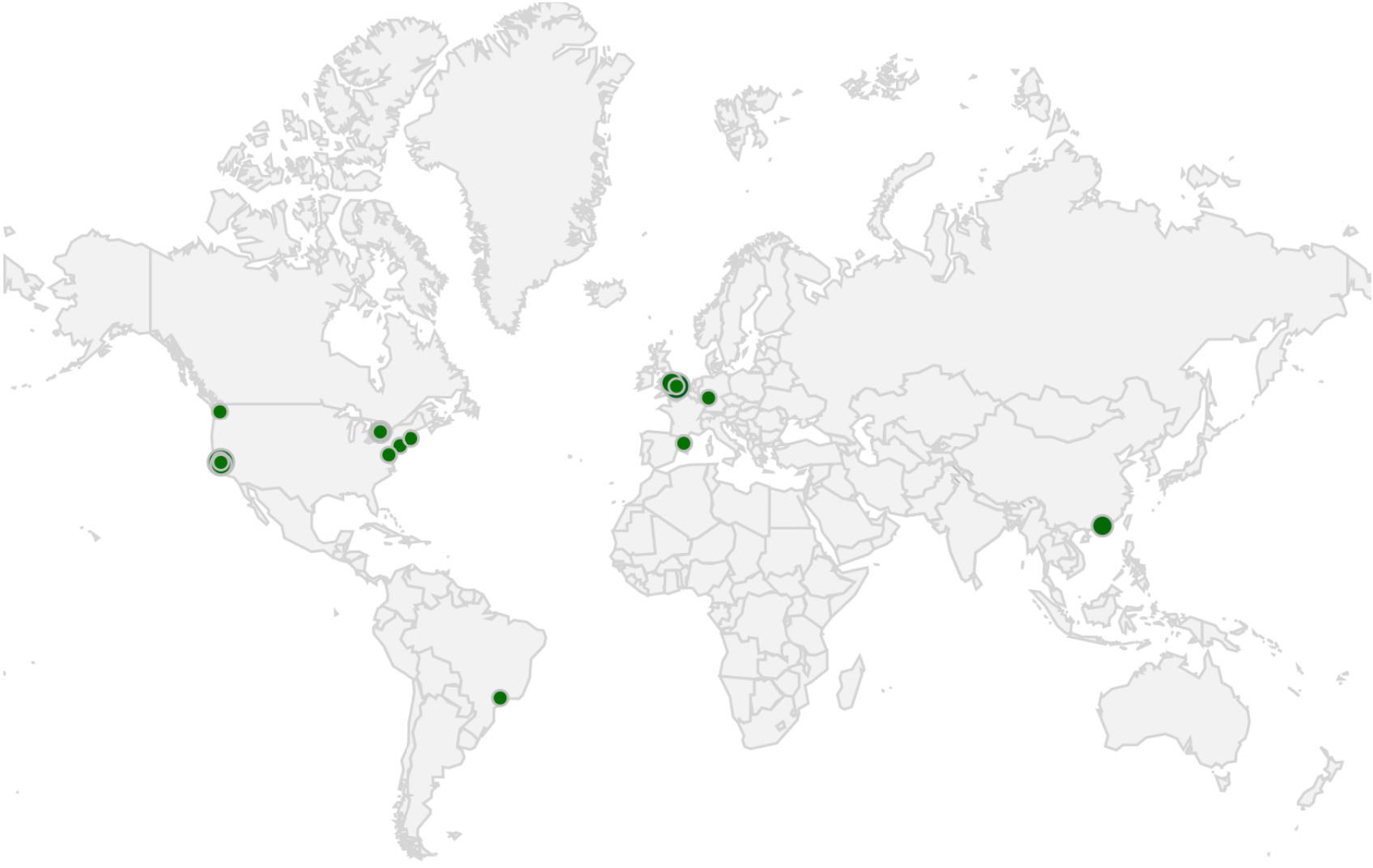
- špecifikácia Beacon API (0.2, 0.3 draft)
- Beacon Development Kits
- monitorovacia infraštruktúra
- Docker obraz
- *klientská aplikácia (nie je súčasťou práce)*

Vyhodnotenie

- veľkosť
- adopcia
- členské organizácie
- dotazy
- užívatelia
- referenčné genómy
- chromozómy
- alely
- škodlivosť mutácií
- vzácnosť variácií
- gény
- ochorenia
- klinické anomálie

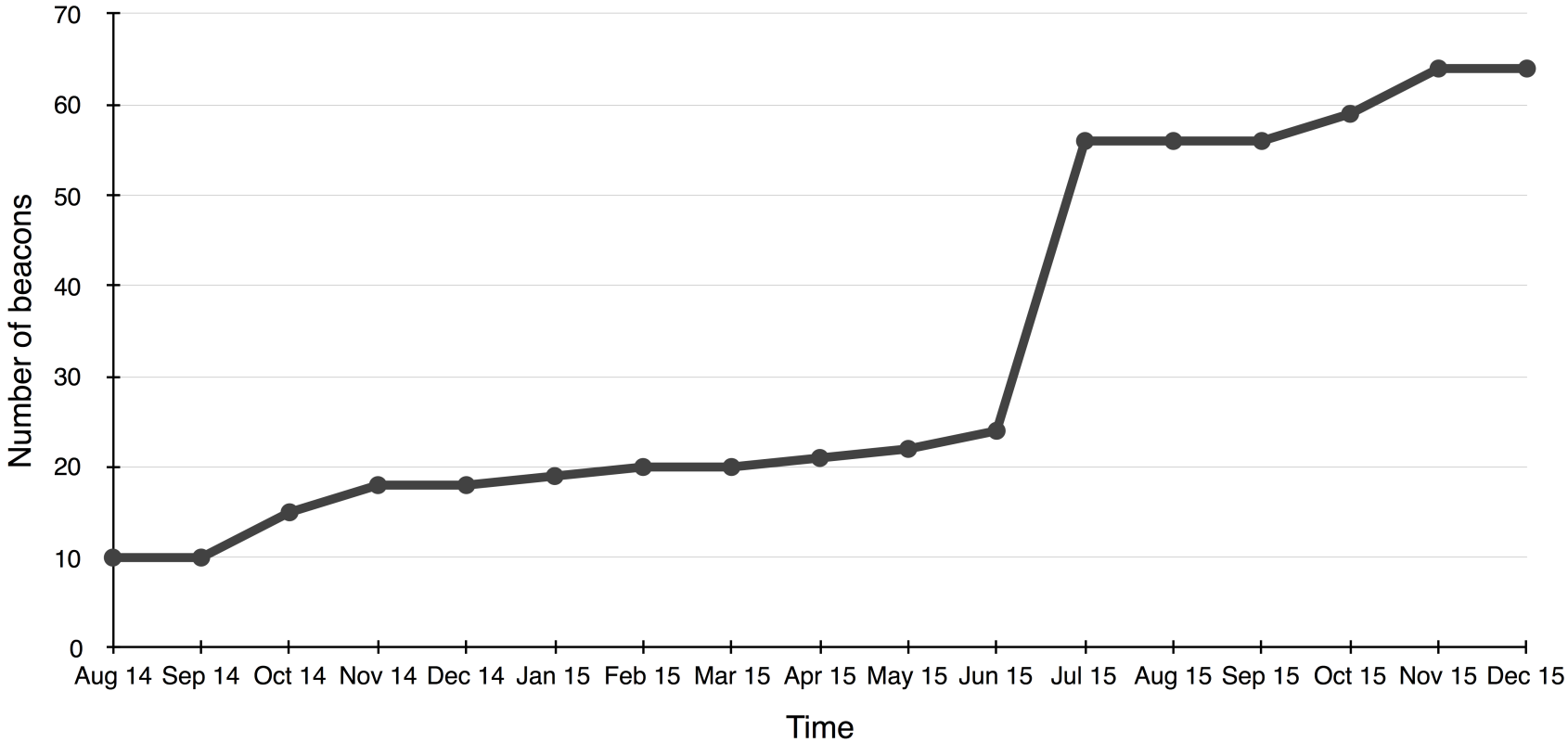
Vyhodnotenie

- globálny systém agregujúci služby zo 7 krajín sveta
- 1 000+ rôznych užívateľov zo 74 krajín



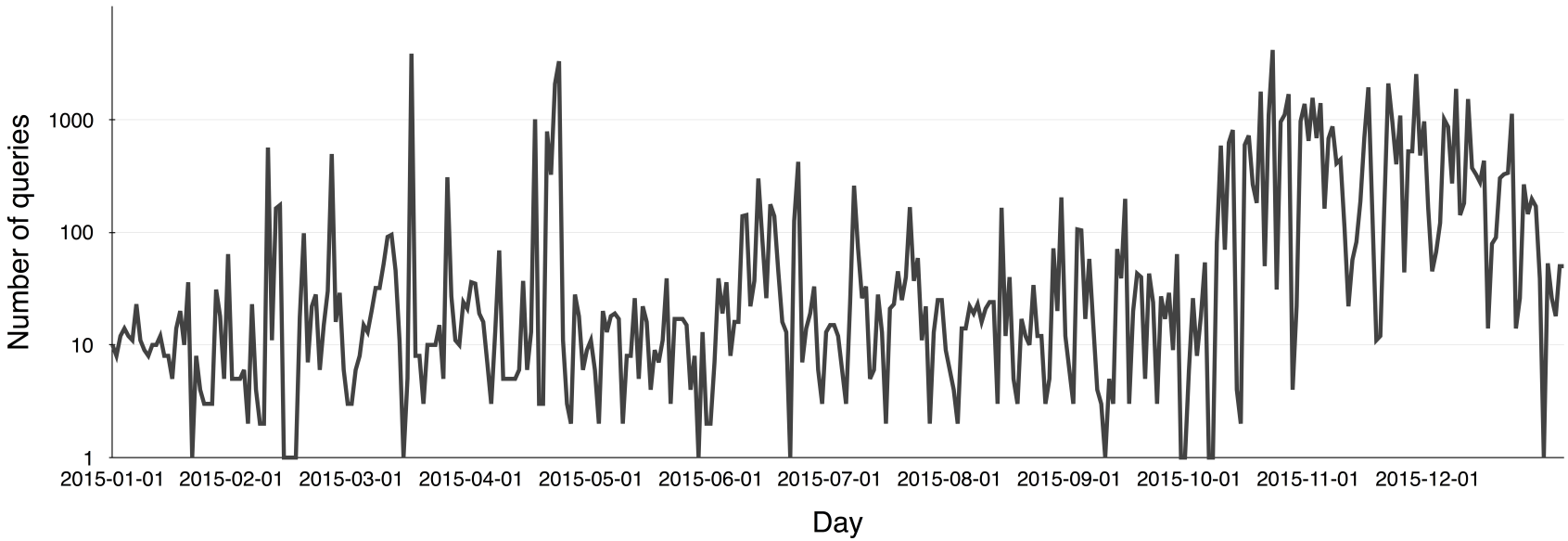
Vyhodnotenie

- 64 Beacon služieb, 23 organizácií
- 160 kolekcií dát, 2 000 000 vzoriek, 2 000 000 000 variácií
- najväčší agregátor genomických variácií človeka na svete



Vyhodnotenie

- 67 000+ prijatých, 800 000+ odoslaných dotazov (2015)
- vysoký počet poškodzujúcich variácií
- netriviálny podiel ojedinelých chorôb
- zaujímavé gény a fenotypy



Zhrnutie

- globálny vzhľadávač genetických dát postavený na protokole Beacon
- úspešný, reálne používaný projekt
- záujem vedeckej komunity
- grant na ďalší vývoj
- predstavované na konferenciách (napr. ASHG, GA4GH Plenary Meeting)
- spomínané v publikáciách – 1 hotová, 1 rozpísaná, 2 plánované
- zmienky v tlači (na webových portáloch)
- používané v reálnych produktoch - Omnicia, Philips (plánované)

Ďakujem.

Otázky?

https://beacon-network.org/

Otázka 1

- **Jaká přesně byla Vaše kontribuce při vytváření Beacon API?**
  Spísanie prvého návrhu 0.2 špecifikácie v JSON formáte pod vedením Beacon týmu, zapracovanie spätnej väzby z diskusie v GA4GH. Návrh 0.3 špecifikácie podľa potrieb Beacon Network, spísanie do Avro IDL, vedenie diskusií o zmenách v API, zapracovanie spätnej väzby, pravidelné aktualizácie o stave na konferenčnom hovore (stále prebieha).
- **Co bylo k dispozici dříve v rámci GA4GH Core API?**
  Protokoly:
  - BEACON – 0.1 (pridané po začiatku tejto práce)
  - GAReads – dátové štruktúry pre základné genetické informácie
  - GACommon, GAReadMethods, GAVariantMethods – pomocné štruktúry a mapovanie na URL

Otázka 2

- **V datovém modelu na Obrázku 3.3 je poměrně nadstandardně velká tabulka Beacon. Jaký k tomu je důvod?**
  Všetky atribúty sú potrebné z niekoľkých dôvodov:
  - id – primárny kľúč
  - organization_id – cudzí kľúč
  - api, auth, description, email, homePage, name, url – štandardné atribúty
  - aggregator, alleleConverter, beaconConverter, chromosomeConverter, enabled, externalUrlParset, fetcher, positionConverter, referenceConverter, requester, responseParser, visible – interné atribúty ovplyvňujúce spracovanie služby
- **Splňuje požadavky základních normálních forem?**
  Áno.