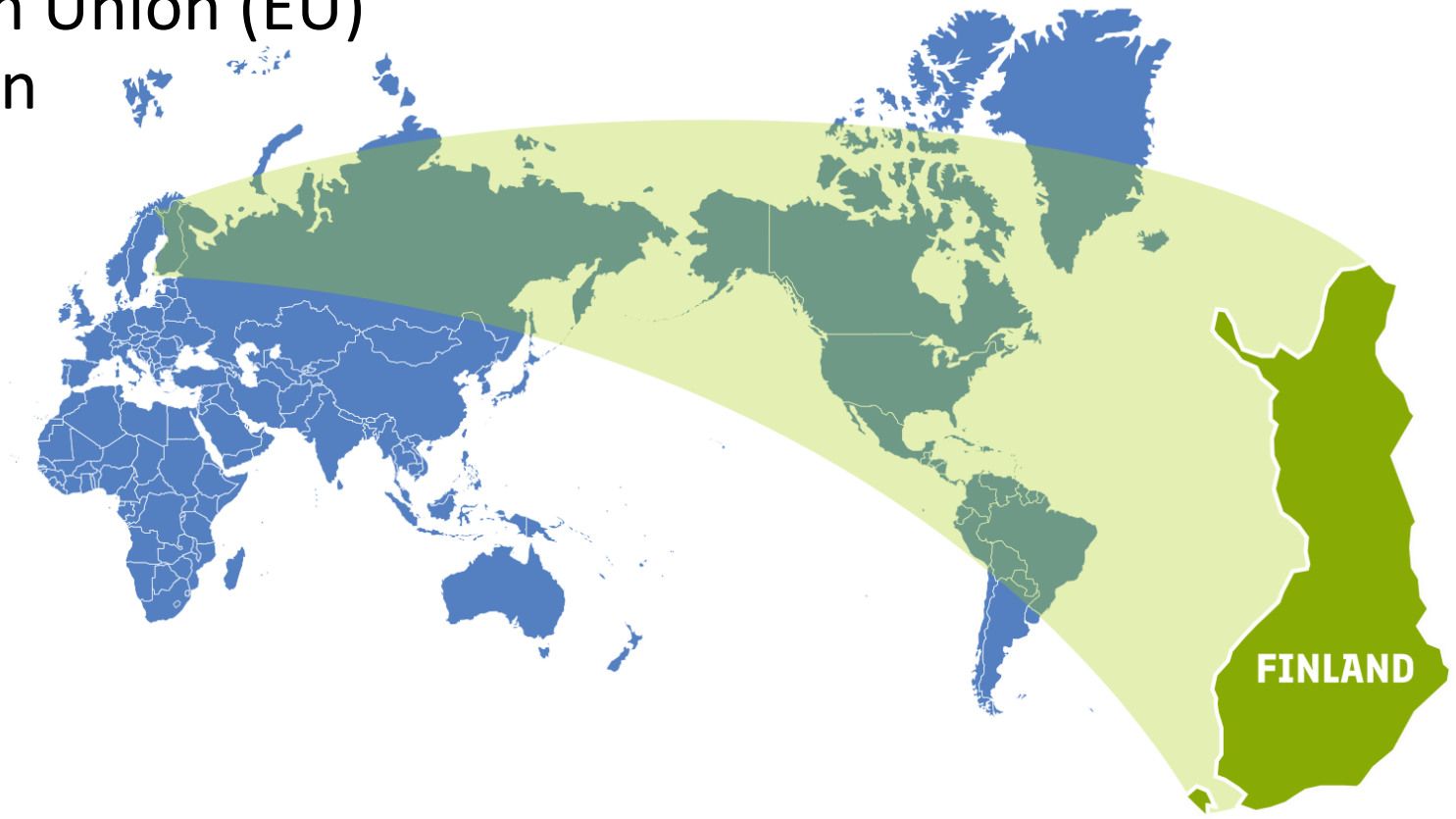
Initial CLDR data collection. A country model from Finland

Erkki I. Kolehmainen and Esko Clarke Sario

Internationalization & Unicode Conference 42, Santa Clara, CA, U.S.A.
12 September 2018

Finland

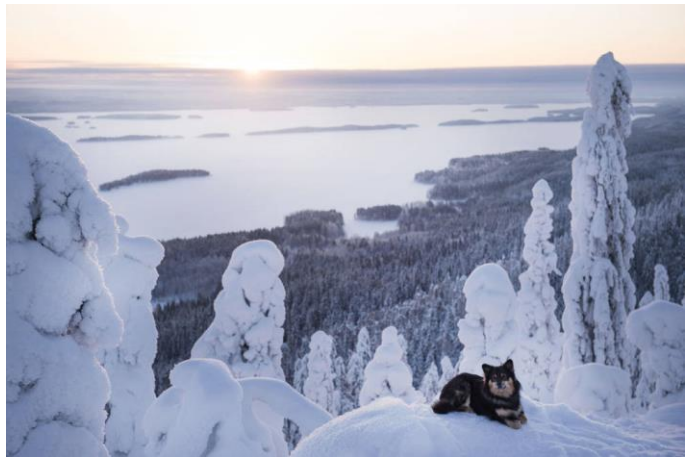
Member of European Union (EU)
Population 5.5 million



Finland

©Julia Kivelä & Visit Finland 2017

©Visit Finland 2009

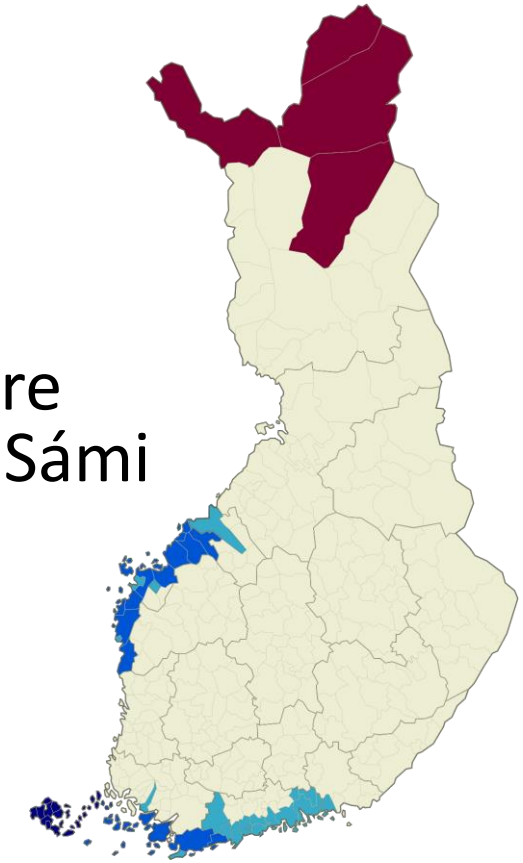


©Jussi Helttunen 2018

©Terhi Ylimäinen 2010

Languages in Finland

- The national languages are
  - Finnish
  - Swedish
- The minority languages with an official status are
  - the Sámi languages: Northern, Inari and Skolt Sámi
  - Romani
  - the Finnish and Swedish sign languages



- Overall, more than 150 languages are spoken in Finland

Map by user:Pottier (fi:Kuva:Suomi.karttapohja.2016.svg)
[Public domain], via Wikimedia Commons

Background – getting into CLDR

- In 2003 Finland was unique in issuing a law over domain names
  - initially for .fi
  - presently also for .ax, for Åland Islands
- The law didn’t consider internationalization at all, in spite of the fact that the core alphabet includes such letters as

  å, ä, ö, š, ž
- CLDR was transferred from the OpenI18N organization to Unicode.
- Guest editorial in the top Finnish daily (Helsingin Sanomat) successfully challenging the Ministry of education into active participation in CLDR. – Birth of Kotoistus Initiative.

Kotoistus initiative

- Coordinates localisation data for Finland
- Steering group representing academia, media, and business
- Openness – anybody is welcome to participate
- Coverage among the important stake holders
- Recruitment by personal contact to the appropriate level of management of the companies
- In the beginning, data to CLDR was entered as XML statements. Later on, the Survey Tool was warmly welcomed

Kotoistus process

- The suggestions for localisation data are being proposed by the coordinator and members of the steering group
- Experts are being consulted during the preparation work
- Working groups, according to areas of interest and expertise, and chaired by steering group members
  - languages
  - currencies
  - countries and regions
  - calendar info and presentation rules
  - etc

Kotoistus process

- Proposals were publicly available for anybody to view and comment.
- Registered members were notified of all proposals.
- All comments were published on the site; thus the quality of them was kept high by self-control.
- In the beginning of the work, only the values that reached consensus were approved and published as recommendations.
- CLDR data was populated based on the recommendations.
- In case of disagreements, the proposal was modified and the next round of discussion was started.

Principles

- User centred and usability oriented approach
- Ambition level comprehensive (with pragmatic omissions)

- The data collected and approved for CLDR is based on what is used in reality in Finland for the target language, e.g.,
  - Currency symbols – only the ones that a Finn can be expected to recognise without problems (€, $, £, ¥, ₽ and the old Finnish markka symbol, mk) are being used, otherwise the three letter currency code as symbol
  - For time zones, we promote Universal Time, and don’t use the time zone abbreviations

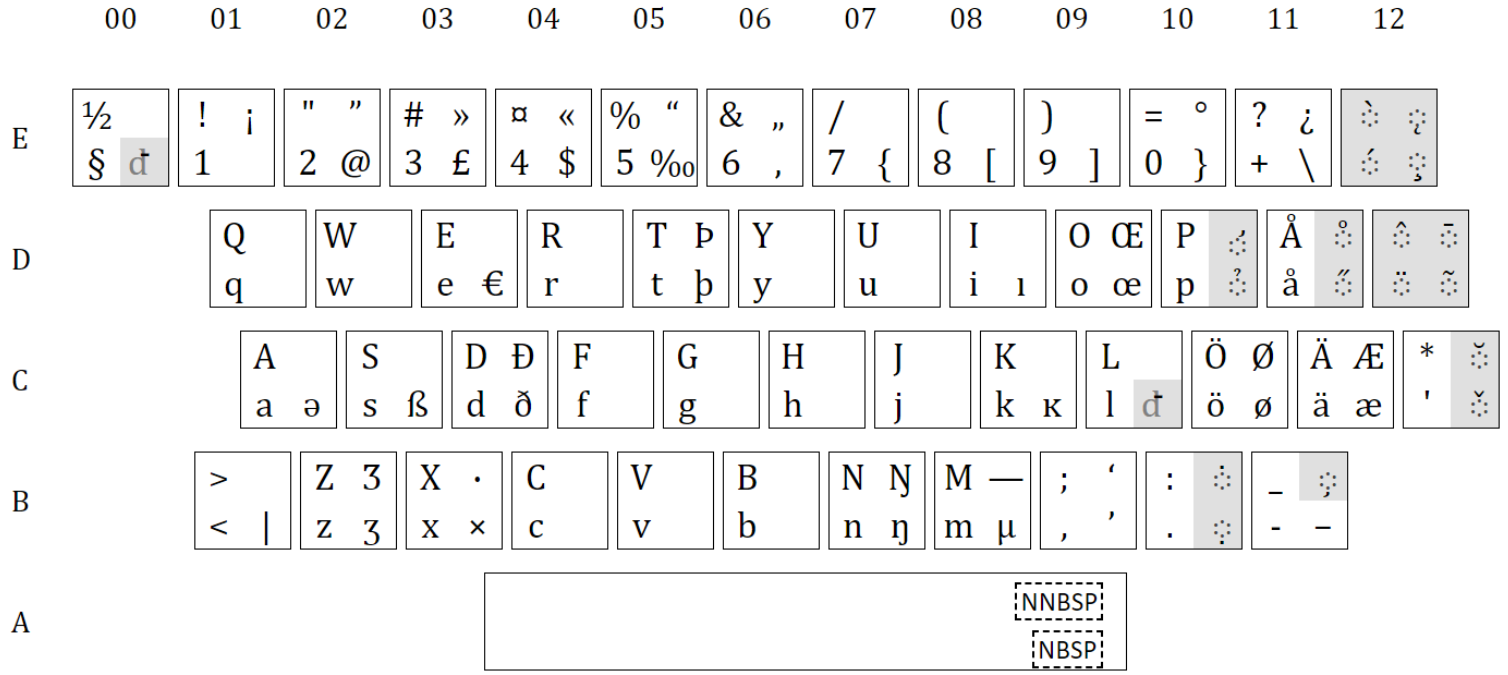
- Fields for currency are in identical order for all currencies – no floating signs

Current process

- Once the bulk of the data has been recorded, the volume of new data is relatively small.
- For most new values, the principles for treating them have been well established.
- Thus, a full-blown consensus seeking would be an overkill, whereas information is still required.
- For new types of data, the original method for consensus seeking is to be used.

Kotoistus, current results

- A national keyboard layout standard (SFS 5966)
  - covering the Latin letters (in current use) of the official EU languages and recognized regional languages, especially those of the Nordic countries (e.g., their Sámi languages)



Kotoistus, current results

- A national ordering standard (SFS-EN 13710)
- Full Uralic Phonetic Alphabet (UPA) in UCS/Unicode
  - over 100 characters
- A number of letters in Landsmålsalfabet
  - those needed for the dictionary of dialects of Swedish in Finland (the rest has still not been added by Sweden)

    ɷ ɐ̯ ɭ
- The list of valid mother tongues been recorded in the Finnish Population Register

Kotoistus, challenges

- CLDR schedules
  - Advance planning and scheduling is a requirement for the process to work, as the input of so many people is needed. Thus, the milestones should be known months in advance. Late changes that bring the deadlines closer effectively hinder success.
  - With the Nordic way of living, mid June to mid August is vacation time, and access to experts is very limited. The same applies from mid December to mid January. Thus a slight move of the CLDR annual circle would benefit us.

Kotoistus, challenges

- Keeping up with how the languages develop while treasuring their uniqueness
  - Spoken language vs. formal language
  - Subcultures, dialects
  - English terminology in technology and research
- Terminology in quick product/update releases
  - Sometimes a term gets established on the market through a product release before the CLDR+Kotoistus process gets a chance to define it for the language

Kotoistus, future

- Focus on the other languages in Finland
  - Swedish, additional work required
  - Northern Sami, additional work required
  - Inari Sami, major additional work required
  - Skolt Sami, initial work required

- Under consideration
  - Romani, initial work required

Finnish model suitability for other locales?

- Most beneficial for the initial collection of data for a locale, or for data within a specific area (e.g. currencies, emoji)
- Requires one finance pool, public and/or private
- Requires commitment from participating organisations.
- The individual experts' voluntary work is essential.
- The cultural climate is open for private + public + commercial co-operation
- Long term commitment by the host necessary for consistent quality

Considerations when starting

- Coordination hosted by a neutral body
  - If not possible, a commercial host that is respected by all participants, and in turn, respects the local experts and language authorities.
- CLDR target level
  - ”modern” to start with
  - yet, while working in the Survey Tool, ”comprehensive” setting should be used to show additional items of interest in the region/language (e.g., former currencies of the country)
- Define the target style of the language
  - User interface language for software and services should suit the users. At the same time, it needs to respect the rules of the language. The need of a healthy future for the language should also be considered. This applies especially on vocabulary, spelling, and use of codes.

Thank you!