The need for public data centers

LDC 20th Anniversary Workshop
The Future of Language Resources

Philadelphia, September 6-7, 2012

Edouard Geoffrois
Direction Générale de l'Armement (DGA)
Agence Nationale de la Recherche (ANR)

Outline

- Why do we need public data centers?
- Where do we stand?
- What are the specific needs in the short and long terms?

The rationale for public data centers

Research in HLT → Data

The rationale for public data centers

Research in HLT → Data

Efficient production and distribution of data → Dedicated centers

The rationale for public data centers

Research in HLT → Data

Efficient production and distribution of data → Dedicated centers

Adequate funding for the needs of public research → Public centers

The rationale for public data centers

Research in HLT → Data ✓ Generally accepted

Efficient production and distribution of data → Dedicated centers ✓ Generally accepted

Adequate funding for the needs of public research → Public centers ✗ Not often discussed

Comparison with the need for evaluation agencies

Research in HLT → Evaluation

Efficient organization of evaluation campaigns → Dedicated agencies

Adequate funding for the needs of public & private research → Public agencies

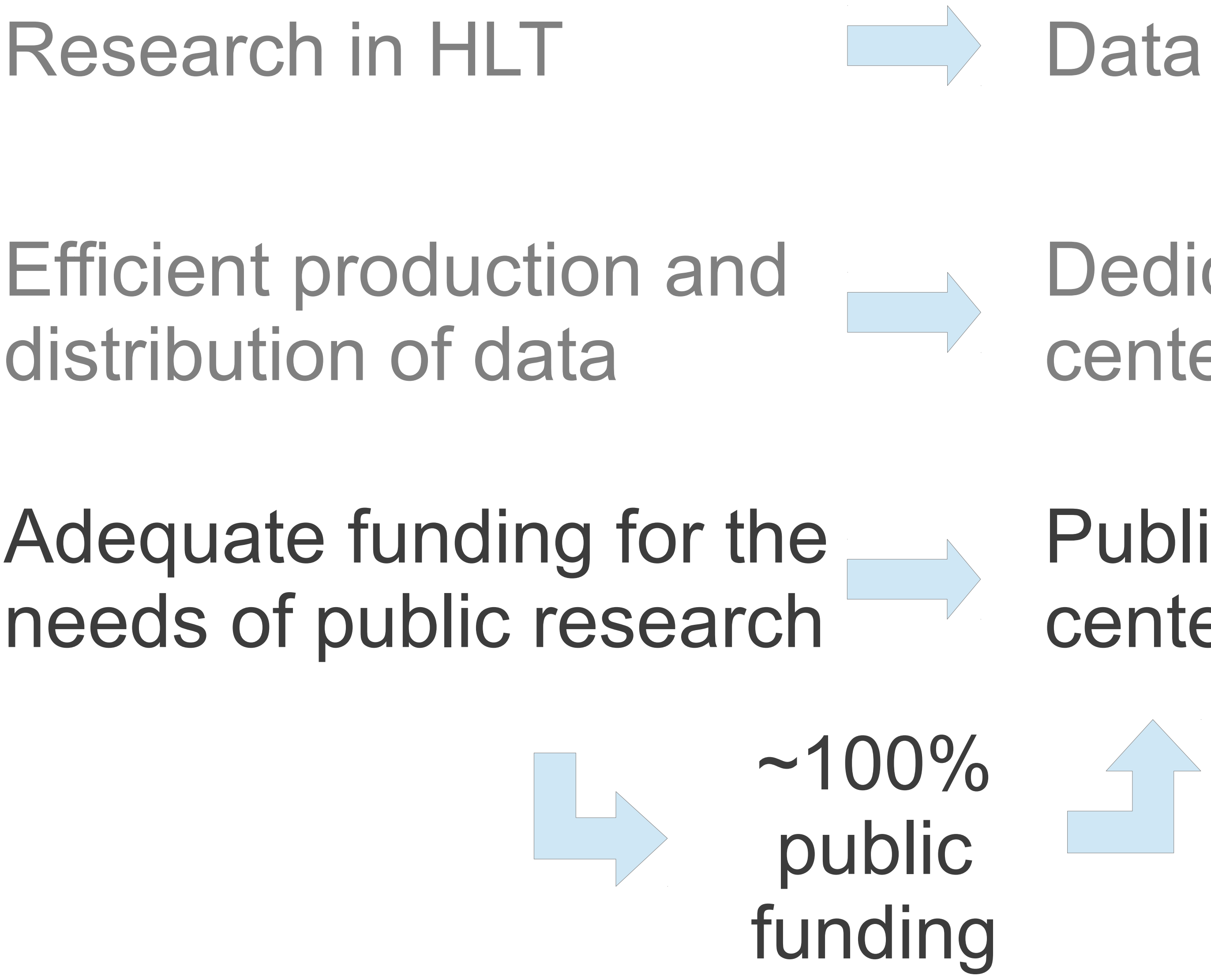
The rationale for public data centers

Research in HLT → Data

Efficient production and distribution of data → Dedicated centers

Adequate funding for the needs of public research → Public centers

The rationale for public data centers



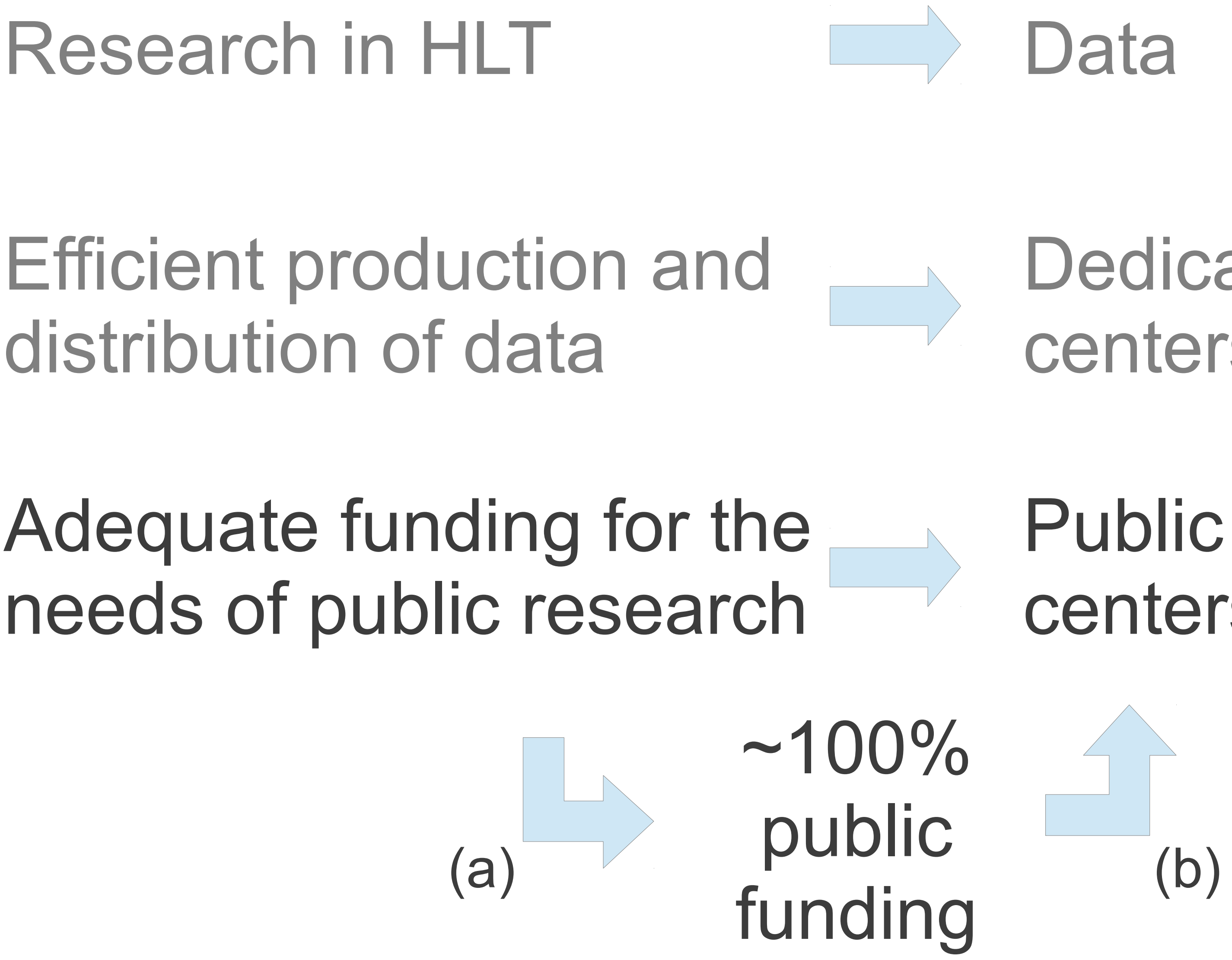
Research in HLT → Data

Efficient production and distribution of data → Dedicated centers

Adequate funding for the needs of public research → Public centers

~100% public funding

The rationale for public data centers



Research in HLT → Data

Efficient production and distribution of data → Dedicated centers

Adequate funding for the needs of public research → Public centers

(a) → ~100% public funding → (b)

Why is it so?

a) Corpora are information goods which are almost public goods (cf. next slide)

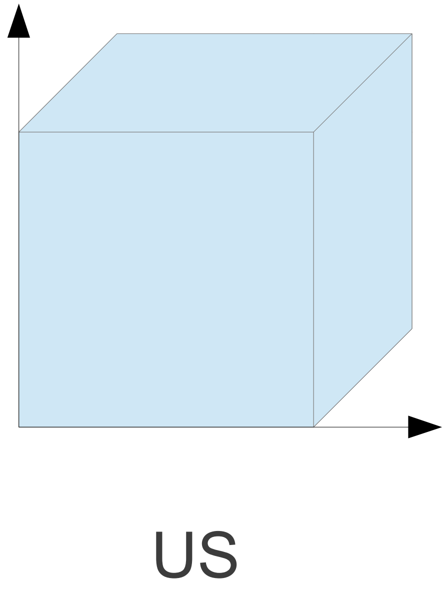
b) There are constraints on public funding of private structures (as a general rule, state aids are forbidden)

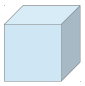
  - Grants can normally provide 100% funding only to not-for-profit organizations
  - Sustainable, long term not-for-profit activities can be achieved only by public recurring funding

Corpora as private/club/public goods

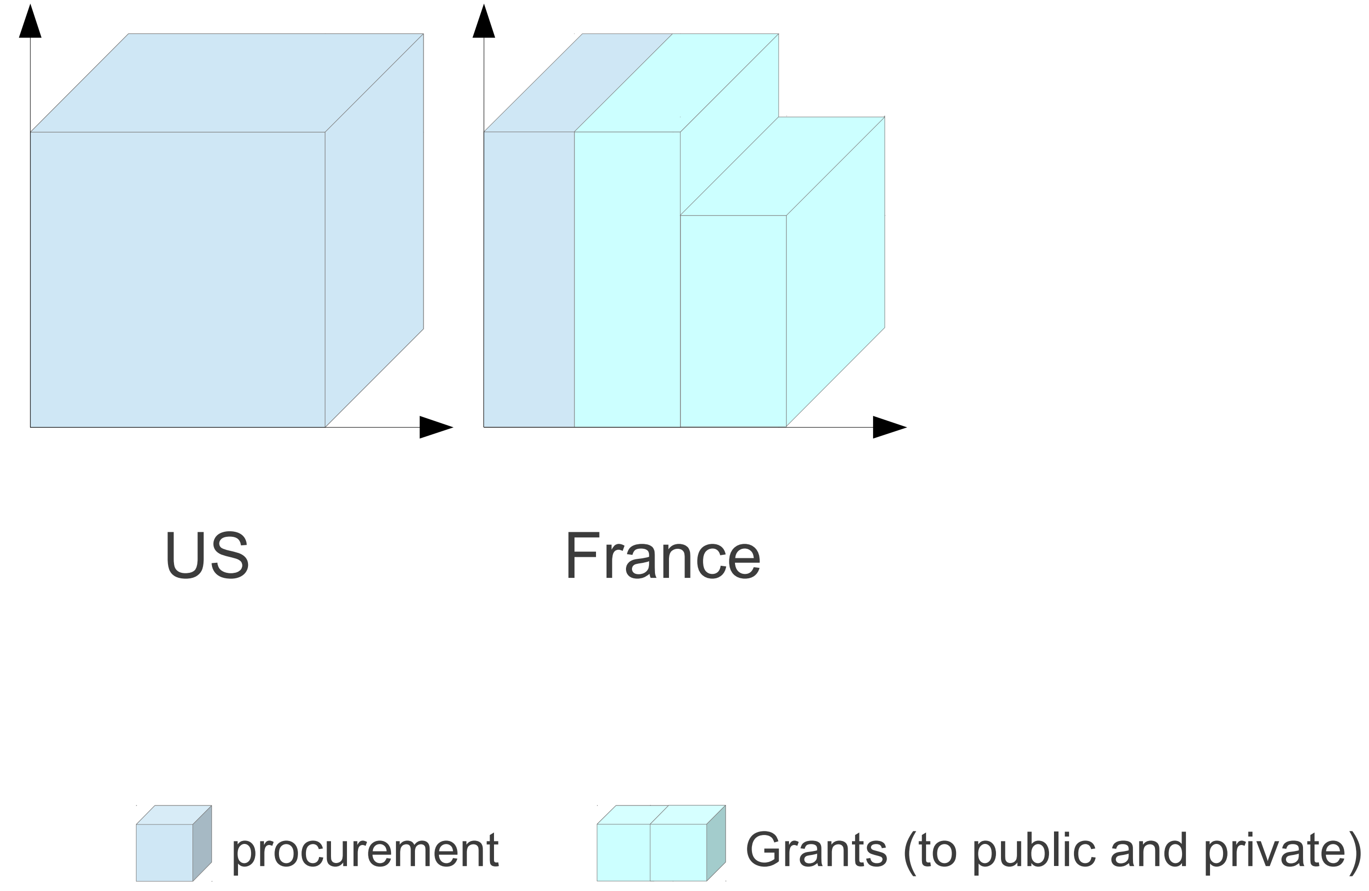
non-excludable	Common goods (e.g., fish stocks, timber, coal)	Public goods (e.g., free-to-air television, air, national defense) Corpus paid by public funding and distributed without a fee
excludable	Private goods (e.g., food, clothing, car, personal electronics) Corpus paid by a company for it own purpose and not distributed	Club goods (e.g., cinema, private parks, satellite television) Corpus sold for a fee
	rivalrous	non-rivalrous

Current situation in the US and Europe

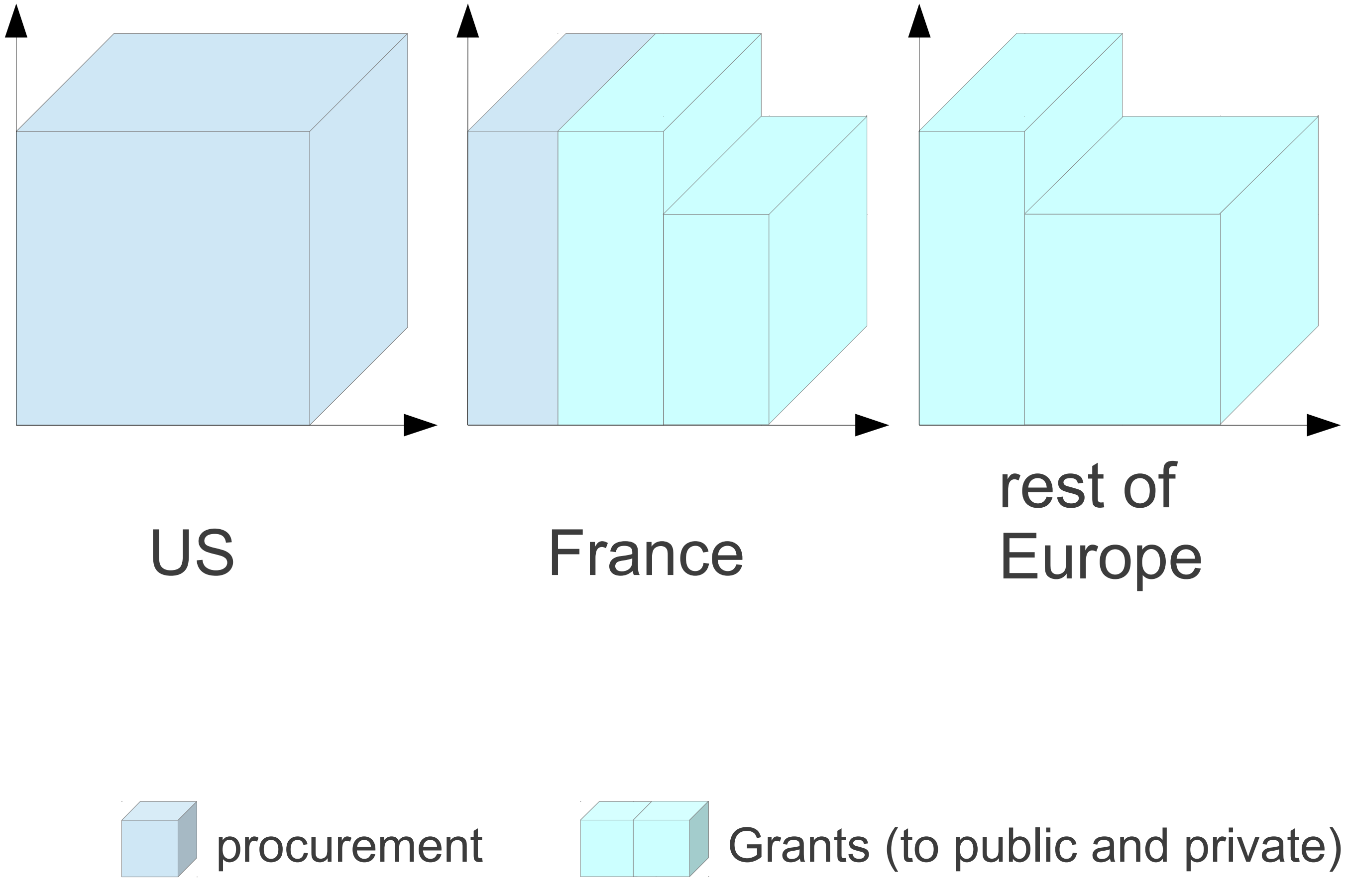


 procurement

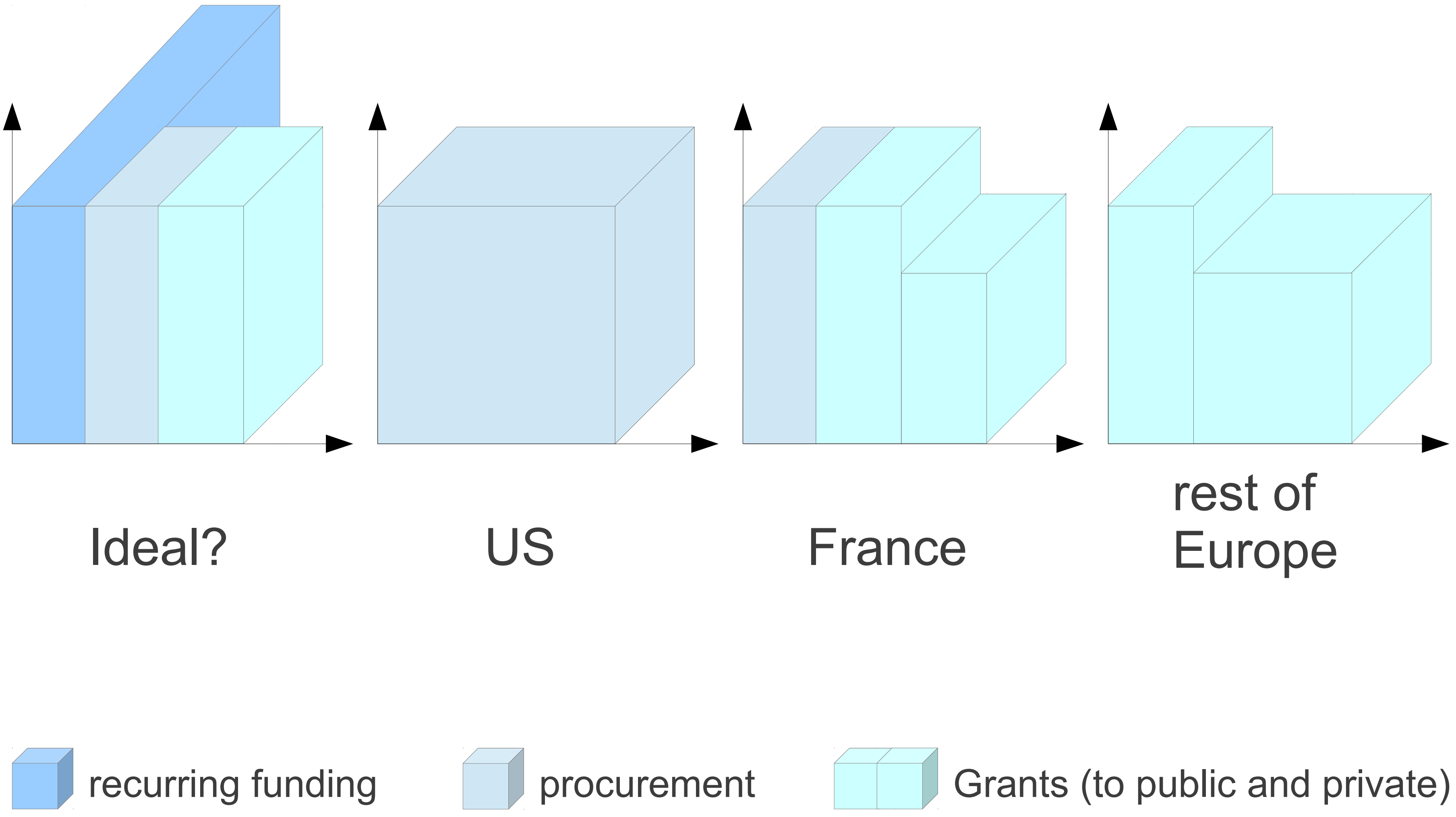
Current situation
in the US and Europe



Current situation in the US and Europe



Current situation in the US and Europe



Ideal?

US

France

rest of Europe

recurring funding

procurement

Grants (to public and private)

Comments

- By using only project-based funding, we bias toward per project needs
  - Some recurring funding would help foster the development and maintenance of generic and reusable data and tools
  - An area for international cooperation
- The line of reasoning applies not only to HLT, but to the whole field of intelligent/multimedia information processing

Specific needs in the short and long terms

- Short term, accepted and emerging
  - More public data centers for HLT in Europe
  - Extension to multimedia information processing
- Short term, but not yet considered
  - Recurrent funding of data centers to foster genericity and reusability
- Longer term
  - A joint international center to develop and maintain reference annotation tools and conventions for sustainable usage of corpora?

Thank you for your attention

and

Happy Birthday to LDC!

Edouard Geoffrois
edouard.geoffrois@m4x.org