When nice people won't share

shy data, web APIs, and beyond

ICGL 2010 (Hong Kong) – Second International Conference on Global Interoperability for Language Resources

today

1. **why nice people don't share**
(people create datasets for different reasons)

2. **from markup to mashup**
(the very nature of sharing has changed)

3. **not just a substitute**
(interoperability is more than interchange)

one: why nice people don’t share

the discourse of

interoperability

has focused on

interchange

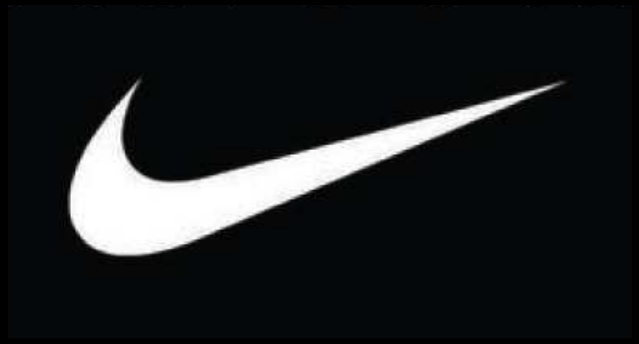
as enabled by

markup

but sometimes, even very nice people,

but sometimes, even very nice people,

WON’T SHARE IT.



they have **shy** data

typically public-funded projects
frequently web-accessible
never instantiatiated
not redistributed

you can meet shy data in public places,
but you can’t take it home with you

the discourse of

interoperability

has focused on

technical issues

the discourse of

interoperability

has focused on

technical issues

but shy data is a **social issue**

in the NEH / US-ED / non-NSF world,
scores of projects are initiated
and managed by

LCTL experts
field linguists
anthropologists
language teachers

data might be

licensed

borrowed

restricted

data might be

licensed

borrowed

restricted

flawed

unfinished

semi-accessible

data might be

licensed

borrowed

restricted

flawed

unfinished

semi-accessible

sui generis

too big to travel

too poor to dress up

Web APIs address the **social** issue
by moving the discourse from

“**Gimme your data.**”

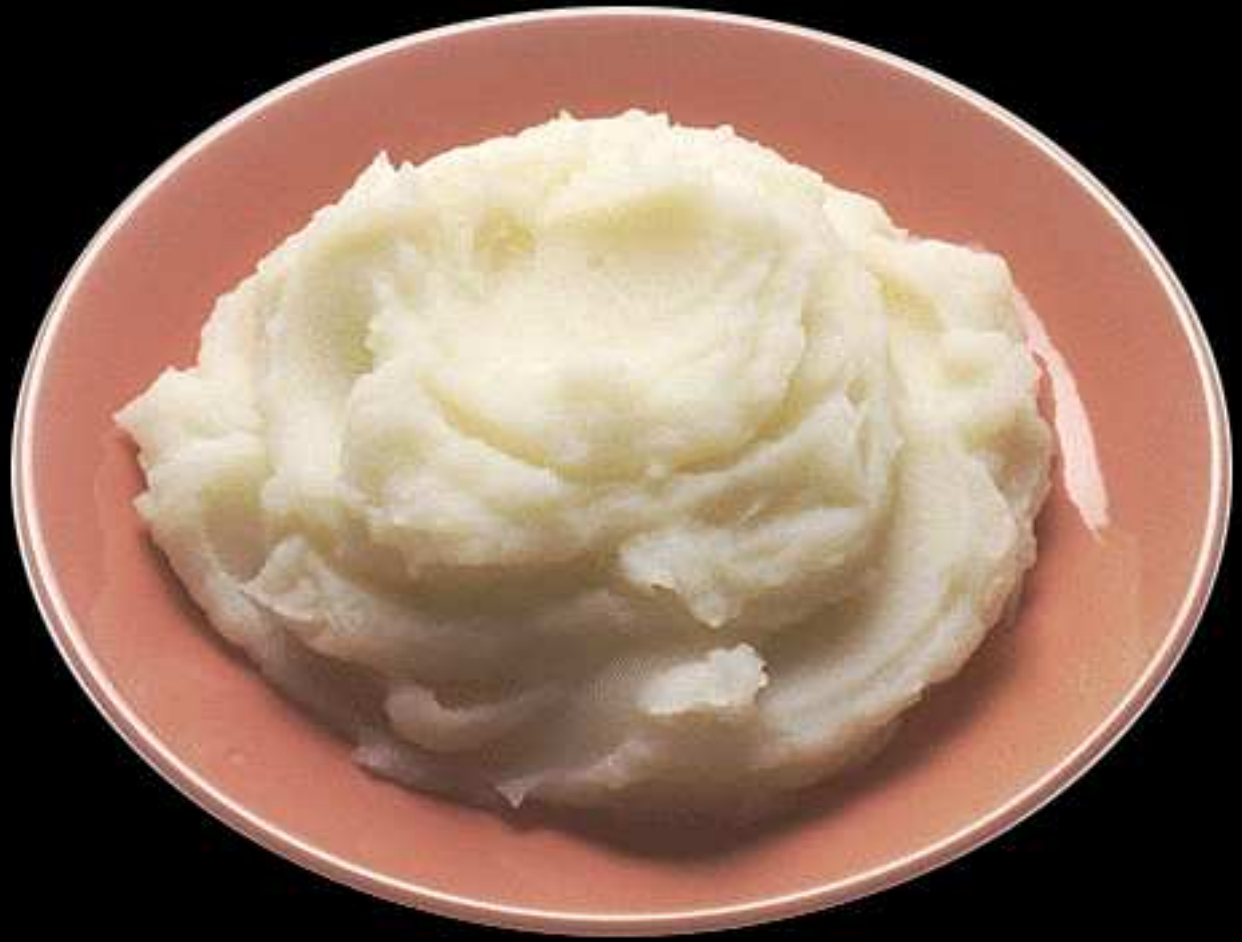
Web APIs address the **social** issue
by moving the discourse from

“**Gimme your data.**”

to

“**May I have this datum, please.**”

two: from **markup** to **mashup**



While we've been busy trading data **sets**, a new market has developed for data **items**.

It's called the Web, and it works pretty much like a restaurant ...

… as long as everything gets
to your table at the same time,

you really don’t care to know
what’s going on in the kitchen.

consider Google

has a **webpage**
has a **database**

but you don’t need to visit
one or download the other

instead, Google exposes an

A P I

that delivers Web services to your plate

when requests for data are

asynchronous
instantaneous
datum-oriented

owning datasets seems … less important

we can get the **milk**

without having to buy (into) the **cow**

APIs take us from warehoused data

to just-in-time data

APIs take us from warehoused data

from numbered releases

to continuous improvement

to just-in-time data

APIs take us from warehoused data
from numbered releases
from heavyweight

to lightweight
to continuous improvement
to just-in-time data

APIs take us from warehoused data
from numbered releases
from heavyweight
from complexity
to simplicity
to lightweight
to continuous improvement
to just-in-time data

API's are to interoperability what modernism was to design

form follows function

developer friendly
inclusive
low-barrier
high yield (vs. metadata farms)
readily back-fit via middleware
graceful degradation / extension
open to continuous improvement
not divorced from services

API downside

persistence persistence persistence
persistence persistence
persistence persistence persistence
persistence
persistence persistence
persistence and persistence

nevertheless

web APIs bring shy data into the game

APIs let you expose data sets

without revealing their content

“On the Internet, nobody knows you’re a dog.”

APIs let you access data sets

without knowing their structure

"I don't <u>need</u> fingers. They have a dog API."

three: not just a **substitute**

the US governent has funded

hundreds

of bilingual dictionaries

why isn’t there an easy

way to look up a

single word

in all of them?

the US governent has funded

hundreds

of bilingual dictionaries

why isn't there an easy

way to look up a

single word

in all of them?

interoperunable

welcome to my world (and projects)

US Dep't of Education / IRS
one resource, many languages
Title VI Interoperability

Language Flagship
one language, many resources
Digital Reference Project

US Dep't of Education / TICFIA
many languages, many resources
Southeast Asian Languages Library

API demo (Russian digital reference project)

Implied call: attribute = *value*

http://russ-flag.org? **service** = *bitext* & **query** = *злой* & **format** = *html* & **number** = *25* & **fold** = *yes* & **resource** = *RNC* & **facet** = *L2:savage* & **lang2** = *EN* & **match** = *both*

BITEXT

Голос был такой громкий и **злой**, что все сразу стихли.

His voice was loud and **savage**, and struck them into silence

API docs

service=collocate

Parameter	Argument(s)	Notes
window	*integer* | *integer,integer* | *integer-integer*	(default **1,1**) Number of words to search on each side of the **query**. *integer* equivalent to **0,1**; return a single table of right-hand collocates *integer,integer* return separate left / right collocate tables *integer-integer* return a single table of collocates
span	**join | split**	(default **join**) Should left-/right-hand neighbors be joined or tabulated separately?
method	**frequency | log | mutual | fisher | tscore | x2**	Method for ranking collocations (built on Ted Pedersen's NSP package; note Fisher is Fisher::left).

Note: returned values include individual frequency, joint frequency,

service=bitext

Parameter	Argument(s)	Notes
facet	*facet:string, …*	A secondary query, generally in English. Known facets are: **L2**: the secondary language.
lang2	**En | …**	The secondary query language (for default **source**)
match	**any | both | L1 | L2**	(default **any**). Allow finer control over the query.
resource	*string*	Name of the corpus (also listed as a common parameter)

Note: multitext resources (like the six-language UN corpus) will use pairwise **resource** *names for now, e.g. RuEnUnCorpos, RuArUnCorpus. In a more generalized definition the ISO 639(-1,2,3) code is incorporated.*

service=dictionary

Parameter	Argument(s)	Notes
facet	*facet:string, …*	Search a particular facet of each entry. There may be multiple **facet** parameters; if so,they are implicitly **and**ed together (all must be met). A typical example: **query=bank&facet=def:robber**

facets reflect **tags**

query: **facets**
return: **tags**

extended **services**

statistics

frequency

lemmatize

headword list

nearest matches

transparent reuse

phoneme distributions

every example sentence

smashups: syns sorted by frequency

first-class **APIs**

validated API

validated API

validated API

validated API

validated API

heterogeneous data

service persistence

service archives

dated **[LIW]AMPs**

link rot

dated respositories

dark archives

in conclusion

people create datasets for different reasons

(1. why nice people don't share)

the very nature of sharing has changed

(2. from markup to mashup)

interoperability is more than interchange

(3. not just a substitute)

When nice people won't share

shy data, web APIs, and beyond

sealang.net/archives/ICGL2010.pdf

ICGL 2010 (Hong Kong)
Second International Conference on Global Interoperability for Language Resources

Doug Cooper

doug.cooper.thailand@gmail.com
Center for Research in Computational Linguistics