CollateX and Interedition

Ronald Haentjens Dekker

Datum: 20-03-2012

Interedition

CollateX

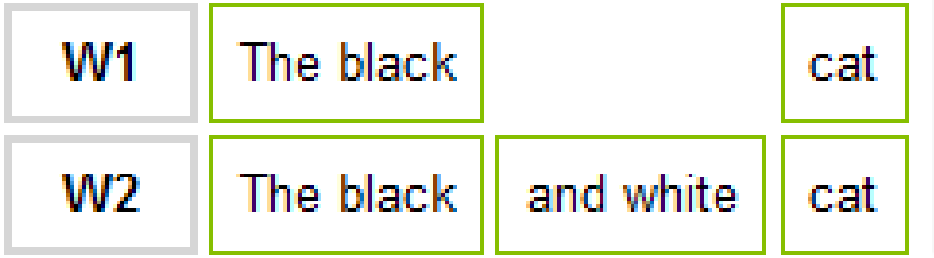
- Prototype for the Interedition COST project
- Interedition is about interoperable tools
- CollateX is a tool: a collation tool
- CollateX is interoperable: it's a service

Collation 1

- Finding differences and similarities between witnesses
- Additions, omissions, modifications
- Compare multiple witnesses against each other
- Not just pairwise comparison
- Order independent

Sequence alignment example

- The black cat
- The black and white cat



Progressive sequence alignment example

- The black cat
- The black and white cat
- The black and green cat

W1	The black			cat
W2	The black	and	white	cat
W3	The black	and	green	cat

Progressive sequence alignment example

- The black cat
- The black and white cat
- The black and green cat
- The black very special cat

W1	The black			cat
W2	The black	and	white	cat
W3	The black	and	green	cat
W4	The black	very special		cat

Progressive sequence alignment example

- The black cat
- The black and white cat
- The black and green cat
- The black very special cat
- The black not very special cat

W1	The black			cat
W2	The black	and	white	cat
W3	The black	and	green	cat
W4	The black		very special	cat
W5	The black	not	very special	cat

Use case

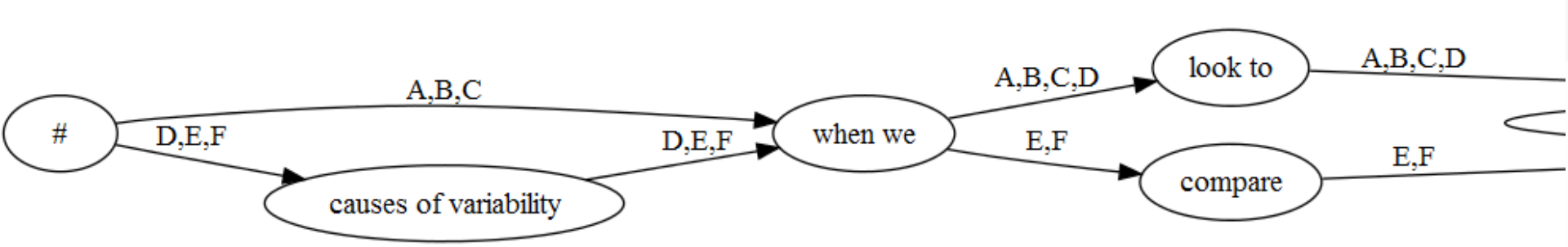
```
<?xml version="1.0" ?>
<examples>
    <example>
        <witness id="1859">WHEN we look to the individuals of the same variety or sub-variety of our older cultivated plants and
            animals, one of the first points which strikes us, is, that they generally differ much more from each other, than do the
            individuals of any one species or variety in a state of nature. When we reflect on the vast diversity of the plants and
            animals which have been cultivated, and which have varied during all ages under the most different climates and
            treatment, I think we are driven to conclude that this greater variability is simply due to our domestic productions
            having been raised under conditions of life not so uniform as, and somewhat different from, those to which the
            parent-species have been exposed under nature. There is, also, I think, some probability in the view propounded by
            Andrew Knight, that this variability may be partly connected with excess of food. It seems pretty clear that organic
            beings must be exposed during several generations to the new conditions of life to cause any appreciable amount of
            variation; and that when the organisation has once begun to vary, it generally continues to vary for many generations.
            No case is on record of a variable being ceasing to be variable under cultivation. Our oldest cultivated plants, such as
            wheat, still often yield new varieties: our oldest domesticated animals are still capable of rapid improvement or
            modification. </witness>
        <witness id="1860">WHEN we look to the individuals of the same variety or sub-variety of our older cultivated plants and
            animals, one of the first points which strikes us, is, that they generally differ more from each other than do the
            individuals of any one species or variety in a state of nature. When we reflect on the vast diversity of the plants and
            animals which have been cultivated, and which have varied during all ages under the most different climates and
            treatment, I think we are driven to conclude that this great variability is simply due to our domestic productions
            having been raised under conditions of life not so uniform as, and somewhat different from, those to which the
            parent-species have been exposed under nature. There is also, I think, some probability in the view propounded by Andrew
            Knight, that this variability may be partly connected with excess of food. It seems pretty clear that organic beings
            must be exposed during several generations to the new conditions of life to cause any appreciable amount of variation;
            and that when the organisation has once begun to vary, it generally continues to vary for many generations. No case is
            on record of a variable being ceasing to be variable under cultivation. Our oldest cultivated plants, such as wheat,
            still often yield new varieties: our oldest domesticated animals are still capable of rapid improvement or modification. </witness>
```

1859	1860	1861	1866	1869	1872
–	–	–	Causes of Variability.	Causes of Variability.	Causes of Variability.
WHEN we	WHEN we	WHEN we	WHEN we	WHEN we	WHEN we
look to	look to	look to	look to	compare	compare
the individuals of the same variety or sub-variety of our older cultivated plants and animals, one of the first points which strikes us, is, that they generally differ	the individuals of the same variety or sub-variety of our older cultivated plants and animals, one of the first points which strikes us, is, that they generally differ	the individuals of the same variety or sub-variety of our older cultivated plants and animals, one of the first points which strikes us, is, that they generally differ	the individuals of the same variety or sub-variety of our older cultivated plants and animals, one of the first points which strikes us, is, that they generally differ	the individuals of the same variety or sub-variety of our older cultivated plants and animals, one of the first points which strikes us is, that they generally differ	the individuals of the same variety or sub-variety of our older cultivated plants and animals, one of the first points which strikes us is, that they generally differ
much	–	–	–	from each other	–
more	more	more	more	more	more
from each other,	from each other	from each other	from each other	–	from each other
than do the individuals of any one species or variety in a state of nature.	than do the individuals of any one species or variety in a state of nature.	than do the individuals of any one species or variety in a state of nature.	than do the individuals of any one species or variety in a state of nature.	than do the individuals of any one species or variety in a state of nature.	than do the individuals of any one species or variety in a state of nature.

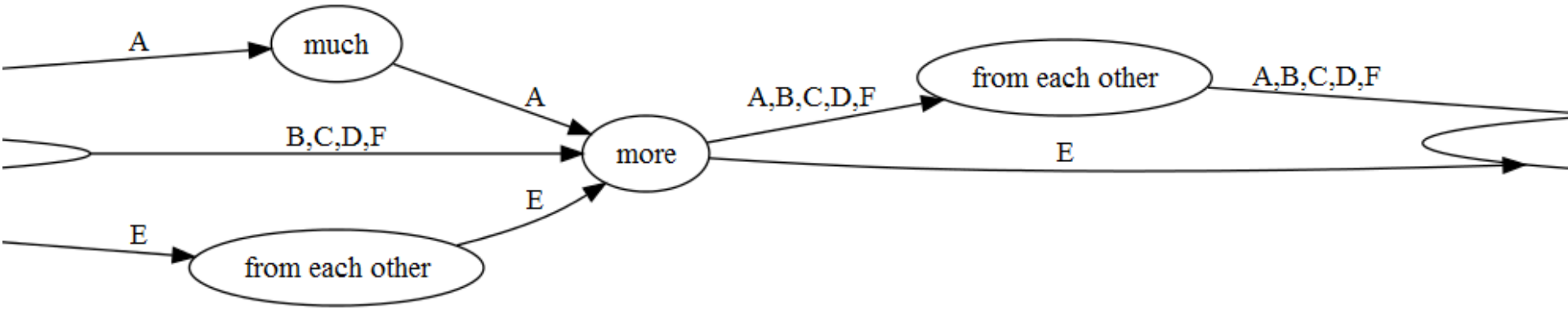
Collation 2

- Finding relations between witnesses
- Parallel segments
- Transposed segments
- Supports multiple output formats
- Can output a variant graph
- Can output TEI Parallel Segmentation alike format

Use case: variant graph



Use case: from each other transposition



Community

- Real life use cases from researchers
- Real life data sets from researchers
- Prototypes tested by researchers and developers
- Build communities of developers through boot camps
- CollateX is open source
- Everyone can fork the source and add to it

Interoperability 1: server

- CollateX is not a desktop application
- It's a service
- Clients connect to the service → workflow, pipeline
- CollateX is platform independent
- It supports Mac OS X, Linux, Windows
- It supports Unicode
- CollateX is a collation framework
- It can support multiple alignment algorithms
- It is extensible: CollateX has an object oriented design

Interoperability 2: clients

- CollateX can be used in many different environments
- On the web: REST API through HTTP, JSON
- In the browser: through javascript
- On the desktop/server: Java as a library
- Cocoon component (XML, XSLT)
- Python bindings
- Exist XML database module?

REST API

```
Content-Type: application/json;charset=UTF-8

{
        "witnesses" : [
                {"id" : "A", "content" : "A black cat in a black basket" },
                {"id" : "B", "content" : "A black cat in a black basket" },
                {"id" : "C", "content" : "A striped cat in a black basket" },
                {"id" : "D", "content" : "A striped cat in a white basket" }
                ]
}
```

```
{"alignment":[
        {"witness":"A","tokens":[
                {"t":"A","n":"a"},
                {"t":"nice","n":"nice"},
                {"t":"black","n":"black"},
                {"t":"cat.","n":"cat"},
                null,
                null,
                null]},
        {"witness":"B","tokens":[
                {"t":"A","n":"a"},
                {"t":"white","n":"white"},
                null,
                {"t":"kitten","n":"cat"},
                {"t":"in","n":"in"},
                {"t":"a","n":"a"},
                {"t":"basket.","n":"basket"}]}
        ]
}
```

Recommendations

- Let researchers and developers work together
- Let developers work together
- Use real life use cases
- Use client-server architecture
- Do only one thing, do it right
- Make tools open source

The end

Interedition