text-fabric

text + annotations
represented for
slicing and dicing - collaboration - versioning

Dirk Roorda
SBL San Antonio
2016-11-18

Data Archiving and Networked Services
DANS

Data Seal of Approval

VU UNIVERSITY AMSTERDAM
EEP TALSTRA CENTRE FOR BIBLE AND COMPUTER

בְּרֵאשִׁית בָּרָא אֱלֹהִים אֵת הַשָּׁמַיִם וְאֵת הָאָרֶץ׃

בְּ	רֵאשִׁית	בָּרָא	אֱלֹהִים	אֵת	הַ
prep	subs	verb	subs	prep	art

בְּ	רֵאשִׁית	בָּרָא	אֱלֹהִים	אֵת	הַ
prep	subs	verb	subs	prep	art
NA NA NA	f sg NA	m sg p3	m pl NA	NA NA NA	NA NA NA

בָּרָא	רֵאשִׁית	בְּ
bārʾā	rēšˌîṯ	bᵊ
בָּרָא ברא]	רֵאשִׁית ראשׁית/	בְּ ב
B.@R@74>	R;>CI73JT	B.:-
B.@R@> BR>[	R;>CIJT R>CJT/	B.:- B
create	beginning	in
none **verb** **verb** hbo 3	none **subs** **subs** hbo 2	none **prep** **prep** hbo 1
m **sg** **p3** NA perf **qal**	**f** **sg** **NA** a NA **NA**	**NA** **NA** **NA** NA NA **NA**
absent absent absent absent absent	n/a absent absent n/a n/a	n/a n/a absent absent n/a n/a
48 746 **15 2341**	***51 707*** **45 868**	***15542 3*** **14194 3**

בָּרָא	רֵאשִׁית	בְּ
bārʾā	rēš ̣îṯ	bᵊ
בָּרָא בר֯א]	רֵאשִׁית ראשׁית/	בְּ ב
B.@R@74>	R;>CI73JT	B.:-
B.@R@> BR>[	R;>CIJT R>CJT/	B.:- B
create	beginning	in
none **verb** verb hbo 3	none **subs** subs hbo 2	none **prep** prep hbo 1
m sg p3 NA perf **qal**	**f sg NA** a NA **NA**	**NA NA NA** NA NA **NA**
absent absent absent absent absent	n/a absent absent n/a n/a	n/a n/a absent absent n/a n/a
48 746 **15 2341**	***51 707*** **45 868**	***15542 3*** **14194 3**
NA Pred **VP** NA *NA* *2* 2	und Time **PP** NA *NA* *1* 1	und Time **PP** NA *NA* *1* 1
? **xQtX** NA *0* **1** *0* *1* 1	? **xQtX** NA *0* **1** *0* *1* 1	? **xQtX** NA *0* **1** *0* *1* 1
1 1	*1* 1	*1* 1

וַ	יַּעַשׂ	אֱלֹהִים	אֶת־	הָ	רָקִיעַ	וַ
and	make	god(s)	<object marker>	the	firmament	and
Conj 65 1	Pred 62 2	Subj 63 3	Objc 64 4	Objc 64 4	Objc 64 4	Conj 65 1
NA 20 1	NA 19 1	NA 19 1	NA 19 1	NA 19 1	NA 19 1	NA 20 1

7

יַּבְדֵּל	בֵּין	הַ	מַּיִם	אֲשֶׁר	מִ	תַּחַת
separate	interval	the	water	<relative>	from	under part
Pred 66 2	Cmpl 67 3	Cmpl 67 3	Cmpl 67 3	Rela 68 1	PreC 69 2	PreC 69 2
NA 20 1	NA 20 1	NA 20 1	NA 20 1	Attr 21 2	Attr 21 2	Attr 21 2

לָ	רָקִיעַ		וּ	רָקִיעַ	הַ	בֵּין	הַ	מַּיִם
to	the	the	and	firmament	the	interval	the	water
PreC 69 2	PreC 69 2	PreC 69 2	Conj 70 4	PreC 69 2	Cmpl 71 5	Cmpl 71 5	Cmpl 71 5	Cmpl 71 5
Attr 21 2	Attr 21 2	Attr 21 2	NA 22 1	Attr 21 2	NA 22 1	NA 22 1	NA 22 1	NA 22 1

אֲשֶׁר	מִ	עַל	לָ		רָקִיעַ	וַ
<relative>	from	upon	to	the	firmament	and
Rela 72 1	PreC 73 2	PreC 73 2	PreC 73 2	PreC 73 2	PreC 73 2	Conj 74 1
Attr 23 3	Attr 23 3	Attr 23 3	Attr 23 3	Attr 23 3	Attr 23 3	NA 24 1

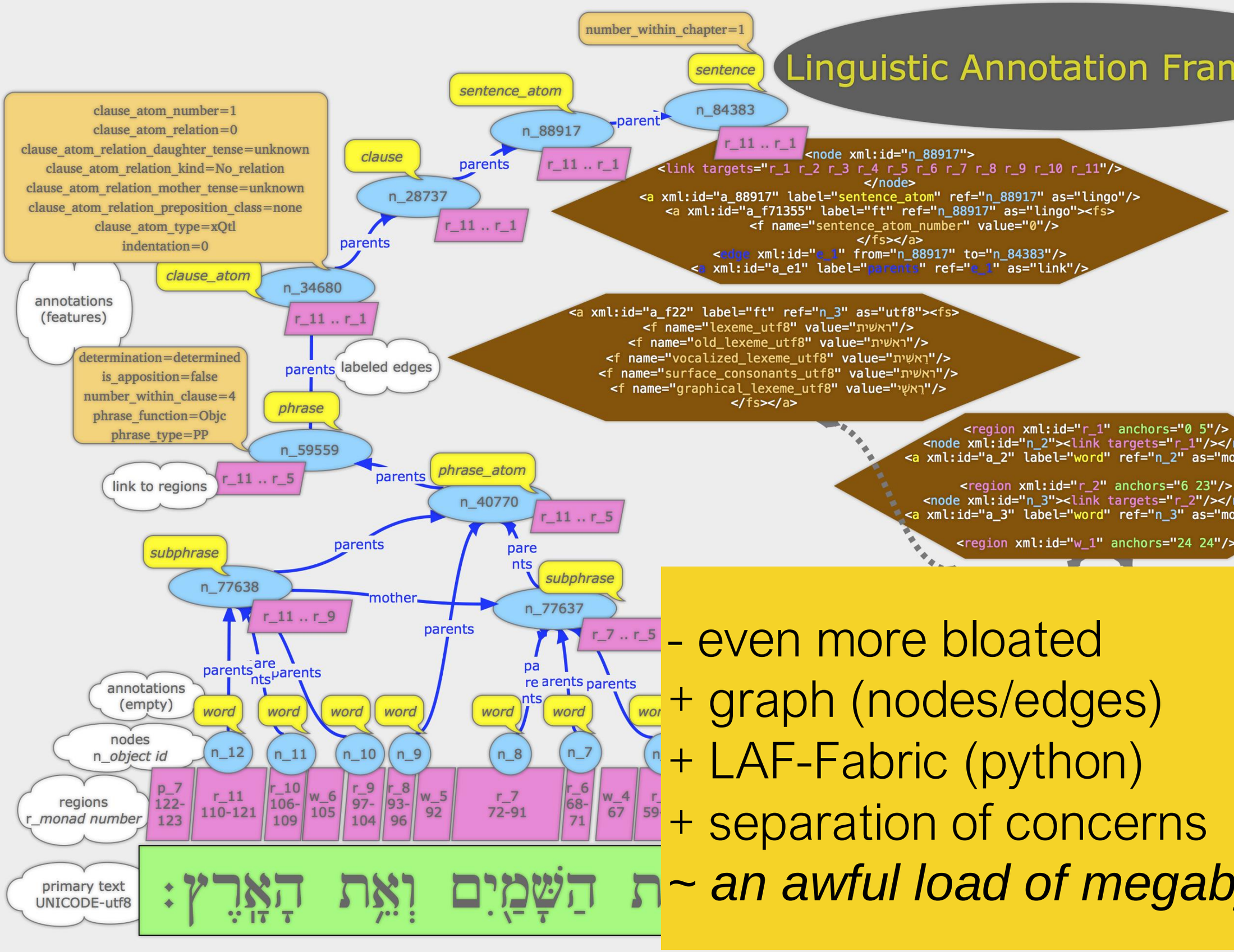
יְהִי־	כֵן׃
be	thus
Pred 75 2	Modi 76 3

straightforward XML

```
...
<clause id="3" rela="Objc">
   <phrase id="27" type="VP">
      <word id="153" sp="verb" lex="BR&gt;[" phono="bārʿā"
        lex_heb="בָּרָא" trans="B.@R@74&gt;"
        gloss="create" partofspeech="verb"
        trailer=" ">בָּרָא</word>
      <word ...>...</word>
      ...
   </phrase>
   <phrase ...>...</phrase>
   ...
</clause>
...
```

- bloated
- hierarchy problems
- inferior tooling (xslt)
- no separation of concerns
- *a horrible mess*

Linguistic Annotation Framework



- even more bloated
+ graph (nodes/edges)
+ LAF-Fabric (python)
+ separation of concerns
~ *an awful load of megabytes*

LAF-Fabric binary

Fn0(etcbc4,ft,language)
Fn0(etcbc4,ft,lex_utf8)
Fn0(etcbc4,ft,lex)
Fn0(etcbc4,ft,ls)
Fn0(etcbc4,ft,mother_object_type)
Fn0(etcbc4,ft,lex)

~ reasonably compact
- a bit opaque
+ column storage
+ python (pickle + gzip)
+ separation of concerns
~ starting to look decent

Fn0(etcbc4,ft,st)
Fn0(etcbc4,ft,tab)
Fn0(etcbc4,ft,trailer_utf8)

```
0: "B"
1: "R>CJT/"
2: "BR>["
3: ">LHJM/"
4: ">T"
5: "H"
6: "CMJM/"
7: "W"
8: ">T"
9: "H"
10: ">RY/"
11: "W"
12: "H"
13: ">RY/"
14: "HJH["
15: "THW/"
16: "W"
17: "BHW/"
18: "W"
19: "XCK/"
```

first 20 words	*from 200000*
B	200000 BN/
R>CJT/	KL/
BR>[	H
>LHJM/	JWM/
>T	B
H	JWM/
CMJM/	PC<[
W	>DWM/
>T	MN
H	TXT/
>RY/	JD/
W	JHWDH/
H	W
>RY/	MLK[
HJH[	<L
THW/	MLK/
W	W
BHW/	<BR[
W	JWRM/
XCK/	Y<JR=/

Text Fabric

+ compact
- transparent
+ column storage
+ pure python
+ separation of concerns

~ with the syntax bureaucracy out of the way, we can start to look into the real problems:

* **slicing and dicing**
* **collaboration**
* **versioning**

Text Fabric (inside)

LAF-Fabric, with the LAF replaced by Text

Inside TF, the LAF-Fabric API just works:

```
for n in NN():
    if F.otype.v(n) == 'verse':
        label = T.passage(n)
        text = T.words(
            L.d('word', n),
            fmt='pf',
        )
        print('{}\t{}\n'.format(
            label, text,
        ))
```

Text Fabric (skeleton)

A TF dataset is a bunch of text files
with at least *otype* and *monads*

node feature
node ⇒ *otype*

otype

```
0-426580  word
426581-514580 clause
514581-605142 clause_atom
605143-858316 phrase
858317-1125831   phrase_atom
1125832-1189401  sentence
1189402-1253740  sentence_atom
1253741-1367532  subphrase
1367533-1367571  book
1367572-1368500  chapter
1368501-1413680  half_verse
1413681-1436893  verse
```

This is the
skeleton:

- the positions
- the containment

of all text objects

monads

```
1367533   1-28762
1367534   28763-52510
1367572   1-673
1367573   674-1167
1413681   1-11
1413682   12-31
1368501   1-4
1368502   5-11
1125832   1-11
1125833   12-18
1189402   1-11
1189403   12-18
426581 1-11
426582 12-18
514581 1-11
514582 12-18
605143 1-2
605144 3
                9-11
```

edge feature
node × node ⇒

Text Fabric (flesh)

information **content**

node features
node ⇒ *otype*

text_full	trailer	p-s-p
בְּ		prep
רֵאשִׁית	_	subs
בָּרָא	_	verb
אֱלֹהִים	_	subs
אֵת	_	prep
הַ		art
שָּׁמַיִם	_	subs
וְ		conj
אֵת	_	prep
הָ		art
אָרֶץ	:_	subs
וְ		conj
הָ		art
אָרֶץ	_	subs
הָיְתָה	_	verb
תֹּהוּ	_	subs
וָ		conj
בֹּהוּ	_	subs
וְ		conj
חֹשֶׁךְ	_	subs

edge feature
node x node ⇒

parent

```
0-1,858317605143
2,858318  605144
3,858319  605145
4-10,858320  605146
426581,1189402  1125832
514581,605143-605146426581
```

Edges (c't'd)

parent

```
0-1,858317605143
2,858318  605144
3,858319  605145
4-10,858320  605146
514581,605143-605146426581
426581,1189402  1125832
```

edge feature
node x node ⇒

1125832 **sentence**

426581 **clause**

605143 605144 605145 605146 **phrase**

858317 858318 858319 858320 **phrase_atom**

0 1 2 3 4 5 6 7 8 9 10 **word**

bᵊrēšˌîṯ bārʾā ʔᵉlōhʾîm ʔˌēṯ haššāmˌayim wᵊʔˌēṯ hāʔʾāreṣ

skeleton | flesh | **slicing 'n dicing**

monads	otype	text_full	trailer	p-s-p	typ
	word	בְּ		prep	
	word	רֵאשִׁית	_	subs	
	word	בָּרָא	_	verb	
	word	אֱלֹהִים	_	subs	
	word	אֵת	_	prep	
	word	הַ		art	
	word	שָׁמַיִם	_	subs	
	word	וְ		conj	
	word	אֵת	_	prep	
	word	הָ		art	
	word	אָרֶץ	:_	subs	
	word	וְ		conj	
	word	הָ		art	
	word	אָרֶץ	_	subs	
	word	הָיְתָה	_	verb	
	word	תֹהוּ	_	subs	
	word	וָ		conj	
	word	בֹּהוּ	_	subs	
	word	וְ		conj	
600000 0-1	phrase	חֹשֶׁךְ	_	subs	
600001 2	phrase				PP
600002 3	phrase				NP
600003 4-10	phrase				VP
400001 0-10	clause				
400002 11-21	clause				-
400003 22-34	clause				Objc
					Resu

vertical: feature selection

horizontal: object selection

both: *modules*

skeleton	flesh		module	collaboration
monads	otype	text_full	p-s-p	strong
	word	בְּ	prep	8675
	word	רֵאשִׁית	subs	7225
	word	בָּרָא	verb	1254 a
	word	אֱלֹהִים	subs	430
	word	אֵת	prep	853
	word	הַ	art	8676
	word	שָּׁמַיִם	subs	8064
	word	וְ	conj	8678
	word	אֵת	prep	853
	word	הָ	art	8676
	word	אָרֶץ	subs	776
	word	וְ	conj	8678
	word	הָ	art	8676
	word	אָרֶץ	subs	776
	word	הָיְתָה	verb	1961
	word	תֹהוּ	subs	8414
	word	וָ	conj	8678
	word	בֹּהוּ	subs	922
	word	וְ	conj	8678
600000 0-1	phrase	חֹשֶׁךְ	subs	
600001 2	phrase			
600002 3	phrase			
600003 4-10	phrase			
400001 0-10	clause			
400002 11-21	clause			
400003 22-34	clause			

module:
feature *strong*
for *words*

dependency:
on the implicit
monad order!

skeleton @20170101		flesh @20170101	**versioning**	module @20161118
monads	otype	text_full	**@20170101⇐@20161118**	strong
	word	בְּ	0 0	8675
	word	רֵאשִׁית	1 1	7225
	word	בָּרָא	2 2	1254 a
	word	אֱלֹהִים	3 3	430
	word	אֵת	4 4	853
	word	הַ	5 5	8676
	word	שָּׁמַיִם	6 6	8064
	word	וְ	7 7	8678
	word	אֵת	8 8	853
9	**word**	**ש**	**10 9**	8676
	word	הָ	11 10	776
	word	אָרֶץ	12 11	8678
	word	וְ	13 12	8676
	word	הָ	14 13	776
	word	אָרֶץ	15 14	1961
	word	הָיְתָה	16 15	8414
	word	תֹהוּ	17 16	8678
	word	וָ	18 17	922
	word	בֹהוּ	19 18	8678
	word	וְ	20 19	2822
600001 0-1	phrase	חֹשֶׁךְ	21 20	
600002 2	phrase		600001 600000	
600003 3	phrase		600002 600001	
600004 4-11	phrase		600003 600002	
400002 0-11	clause		600004 600003	
400003 12-22	clause		400002 400001	
400004 23-35	clause		400003 400002	
			400004 400003	

old module works as is on new data

the end of the beginning

etcbc	Stephen Ku	David van Acker	James Cuénod
flesh @20160101	skeleton		
	@20160101 <= @20170101		
flesh @20170101	skeleton	module **strong** @20161118	
	@20170101 <= @20180101		
flesh @20180101	skeleton		module **accent** @20170202
	@20180101 <= @20190101		
flesh @20190101	skeleton	module **strong** @20171118	module **assoc** @20180303
	@20190101 <= @20200101		

text-fabric

thank ye

dirk.roorda@dans.knaw.nl

shebanq@ancient-data.org

Data Archiving and Networked Services

DANS

Data Seal of Approval

VU UNIVERSITY AMSTERDAM

EEP TALSTRA CENTRE FOR BIBLE AND COMPUTER