How to be Lazily Productive in Evaluation

Olivier Hamon

ELDA

November 27, 2012

Sommaire

Our main issue : task repetition

CLARA

Why "lazily" ?

- Evaluation is nice, but can become really repetitive
- Conversion of the repetitive tasks into generic and sustainable ones

Why "productive" ?

- Evaluation generally at the end of the workflow or cycle
- ... often upstream delays...
- ... often impatient system developers...

- Quick results, but not to the detriment of quality and reliability
- Don't forget the cost ! ⇒ data creation, human judgements, workflow management, etc.

Solutions

CLARA

DRY : Don't repeat yourself !

- Principle in software development
- Can be adapted to evaluation

Don't repeat evaluation but rather reuse :

- The methodology : protocols, workflows, measures...
- The data : in-domain project transfers, system development, cross-domain project transfers...
- The tools : metrics, interfaces, platforms...

(Solutions)

CLARA

Non-sustainable tasks

- Evaluation set up (find data, judges, etc.)
- Quality quantification
- Results analysis and interpretation
- *Although experience helps !*

Context

CLARA

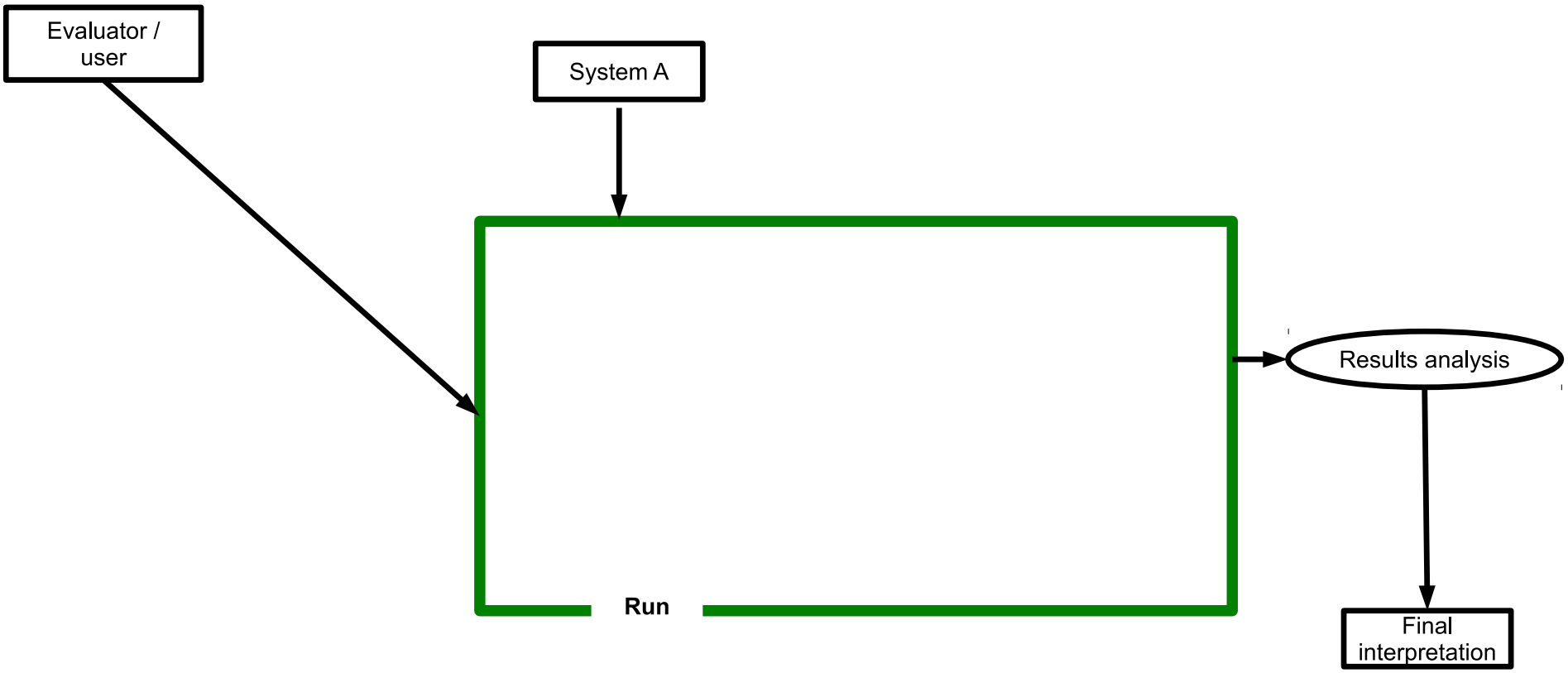
ELDA and the evaluation

- Organisation and collaborations in many evaluation campaigns
- Large volume of experimentations
- Metric development and result analysis
- First evaluation platform implementations

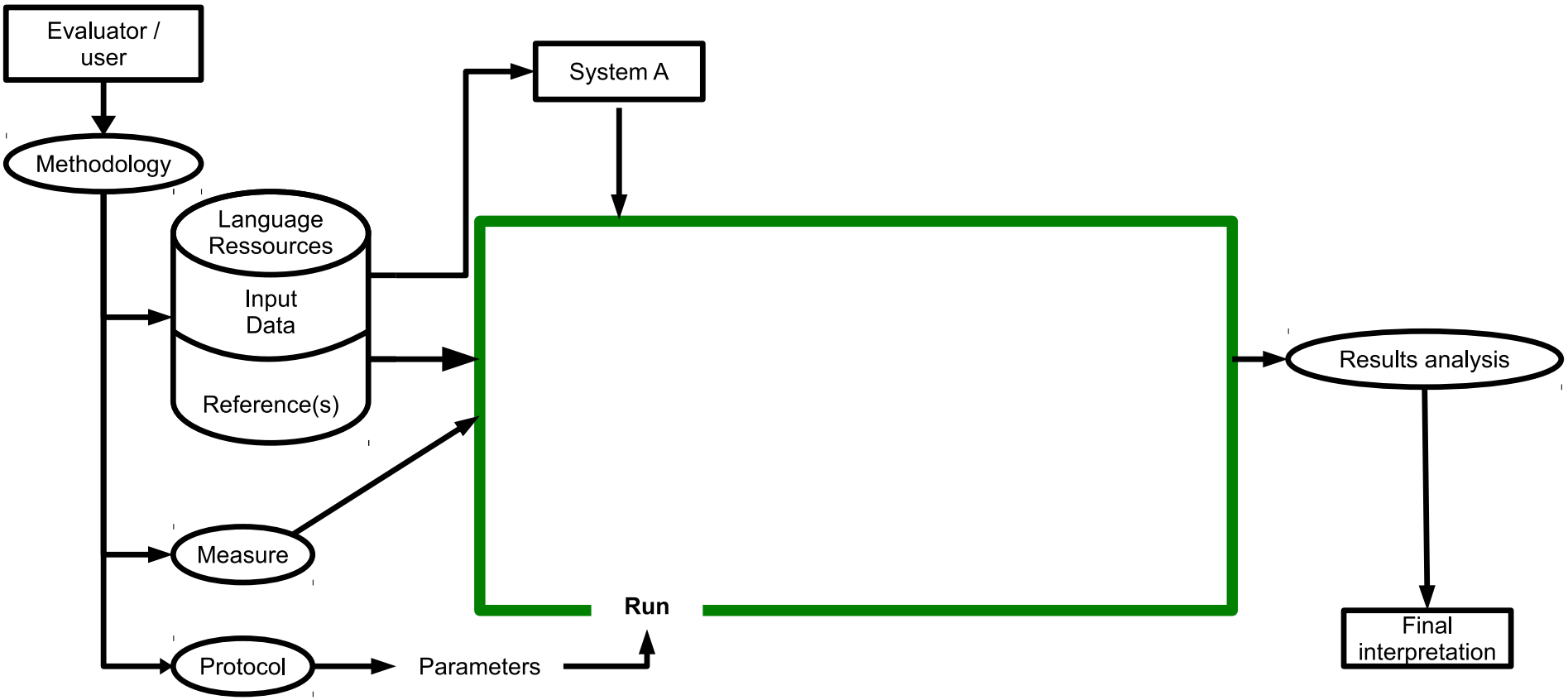
(Still) Growing need for evaluation

- Well established in National and European projects
- System development
- *Often on similar topics and domains !*

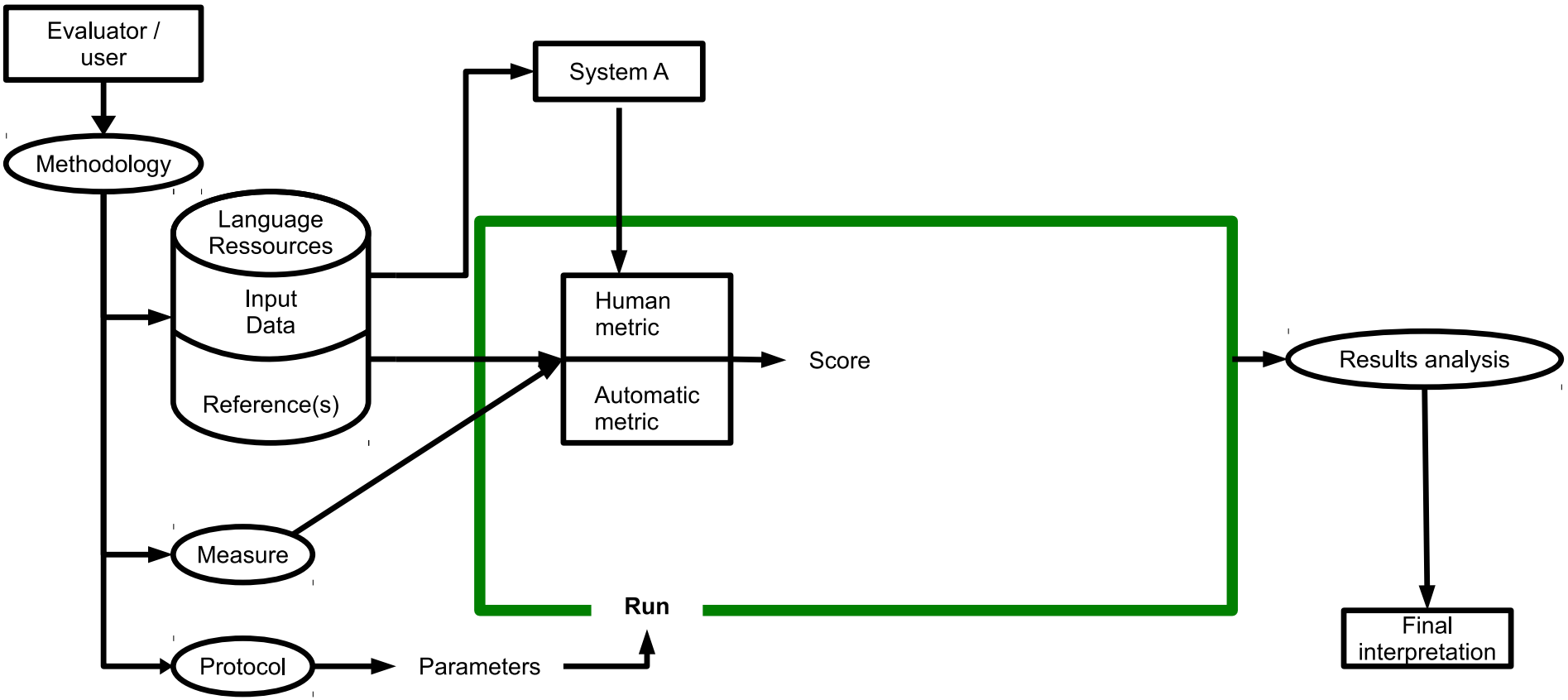
Evaluation Workflow



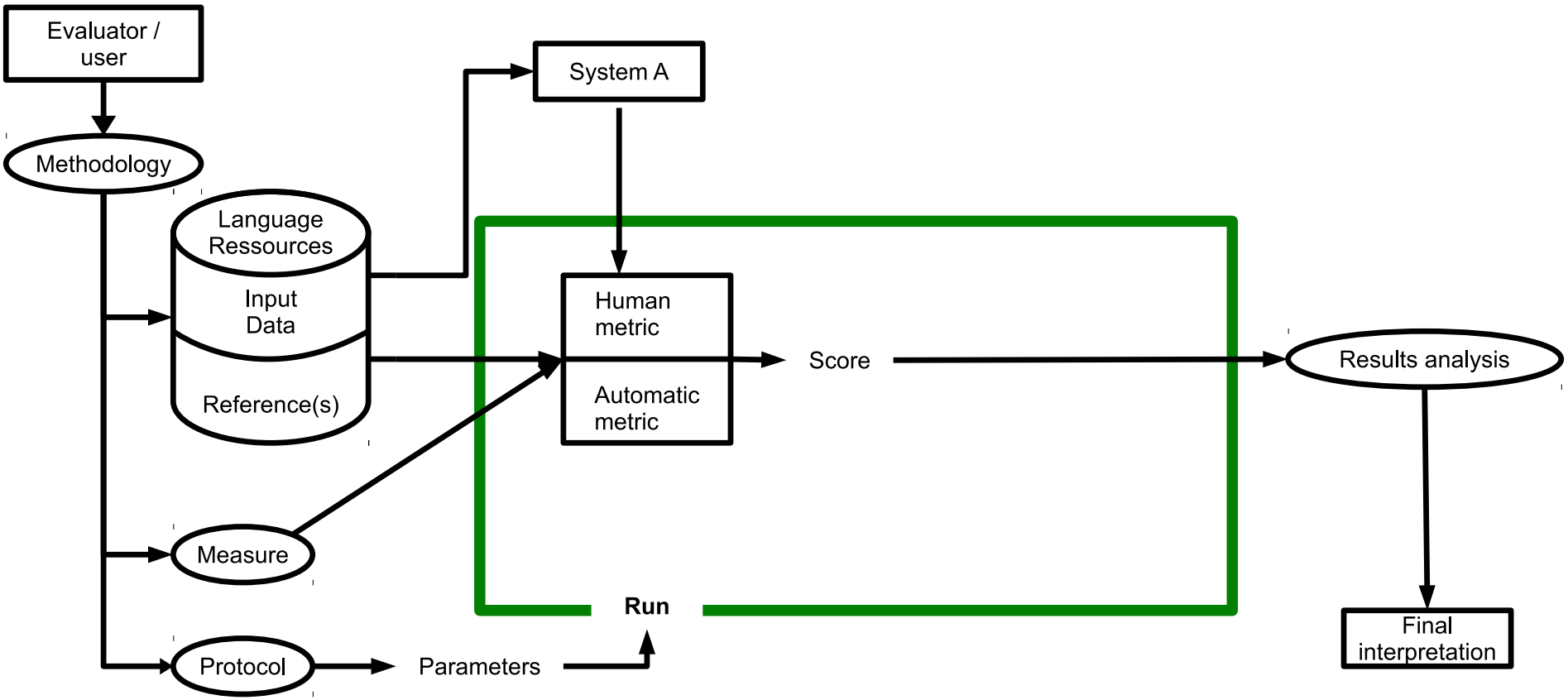
Evaluation Workflow



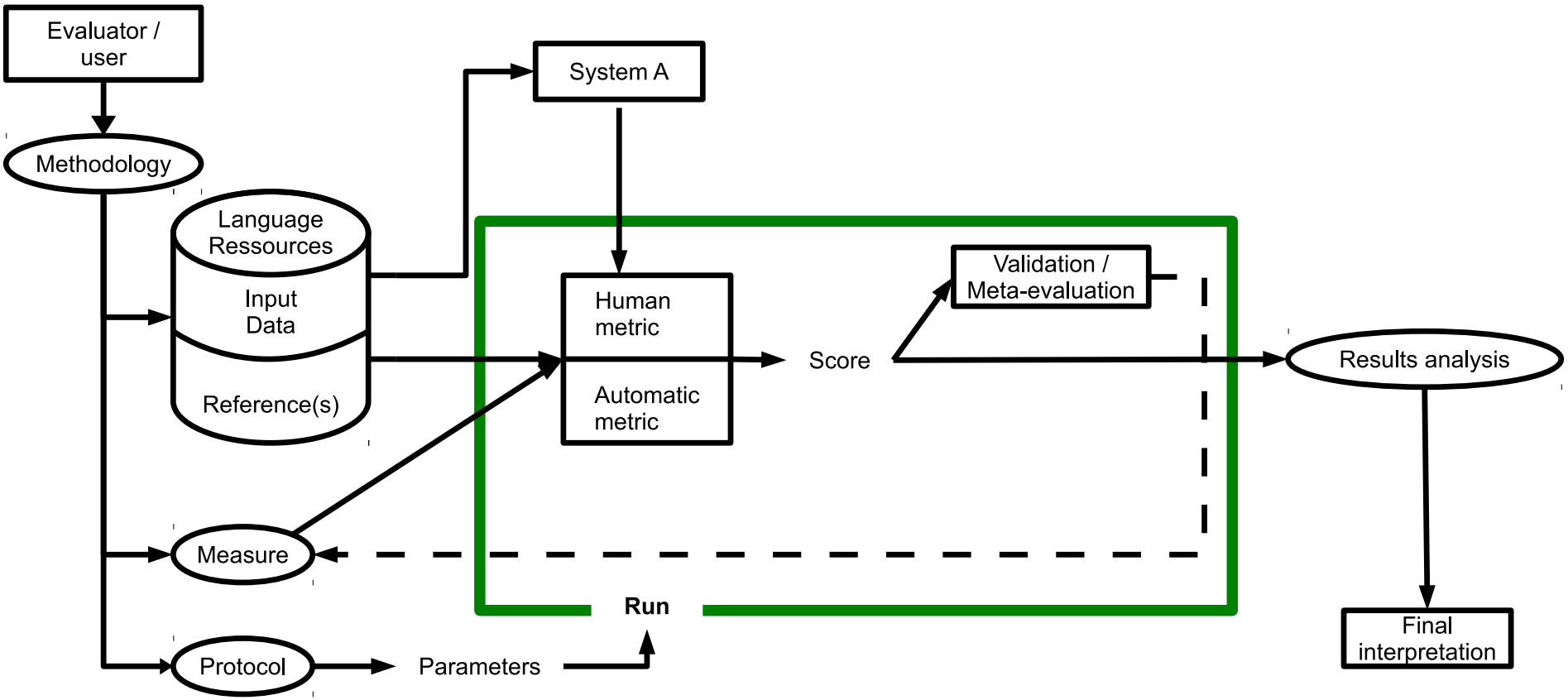
Evaluation Workflow



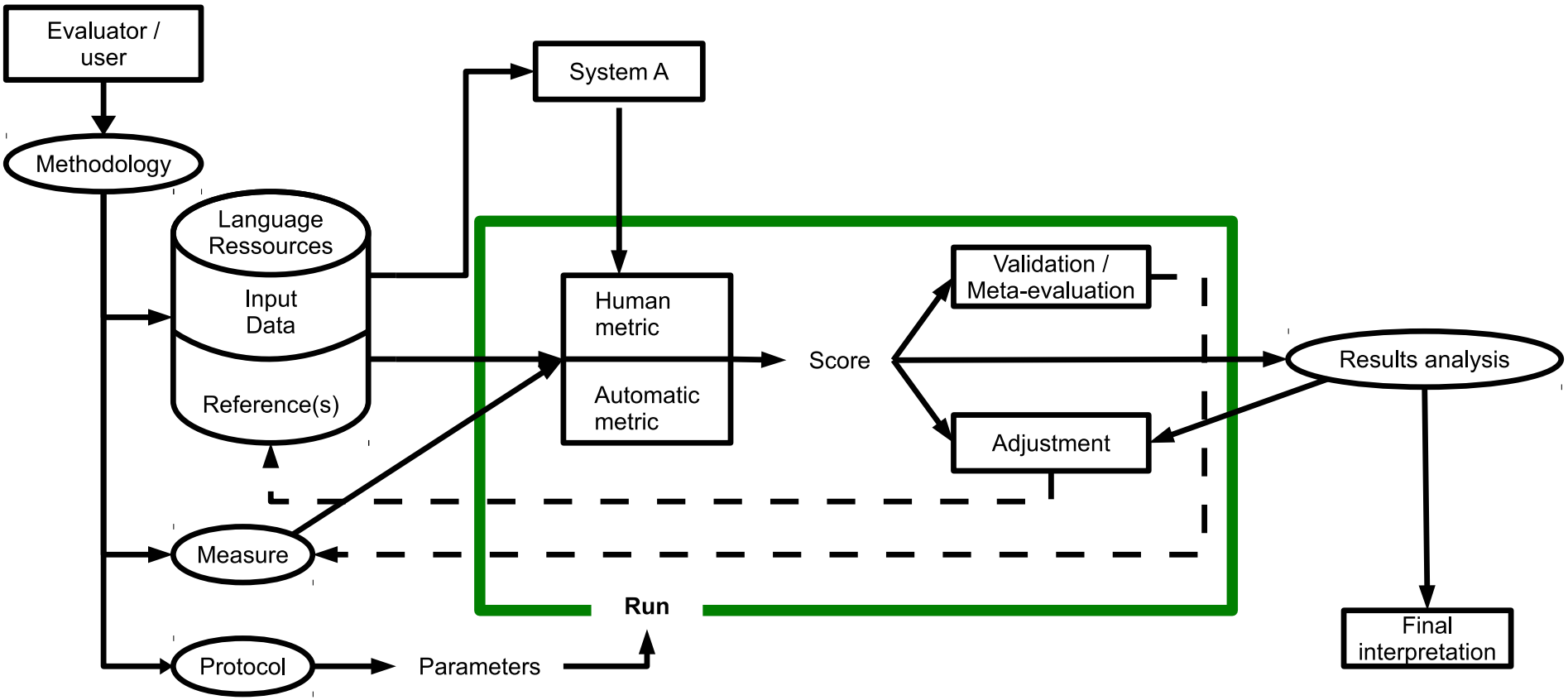
Evaluation Workflow



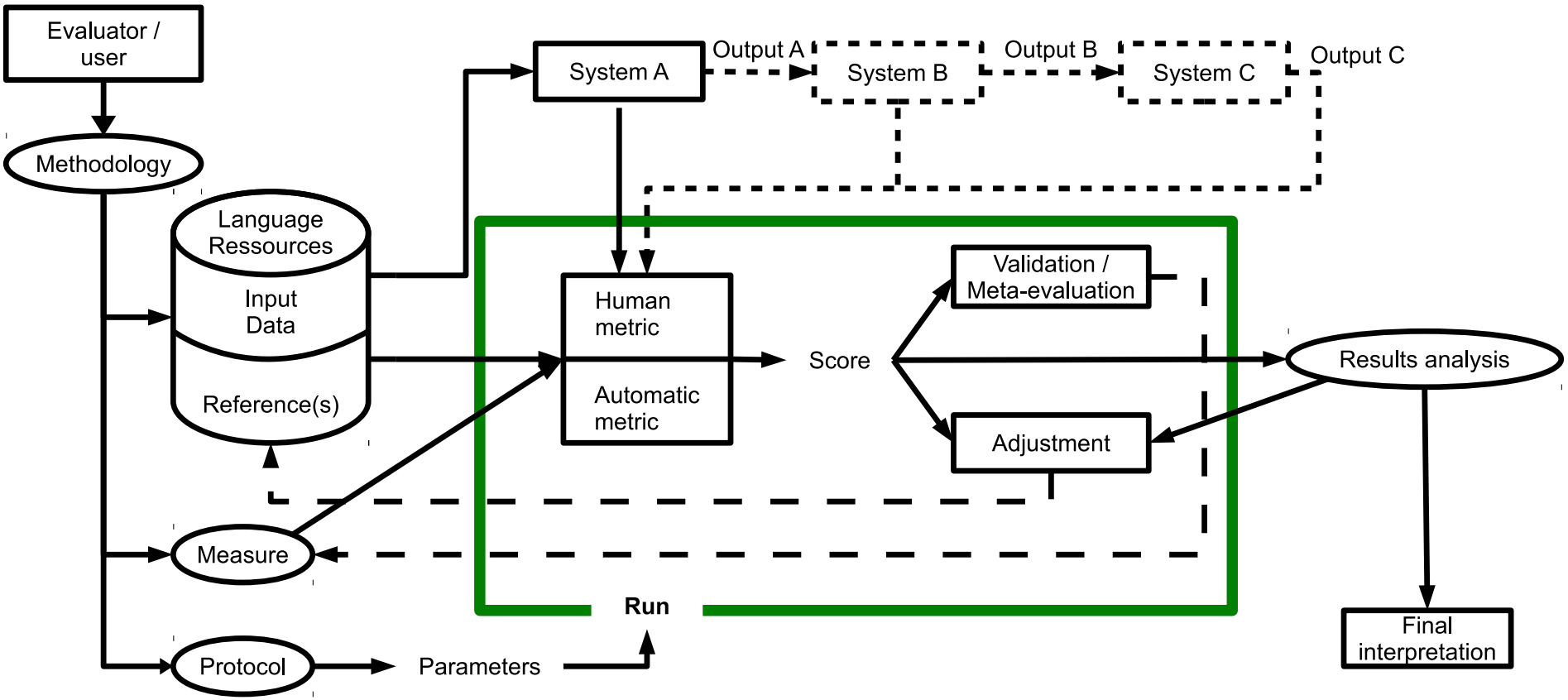
Evaluation Workflow



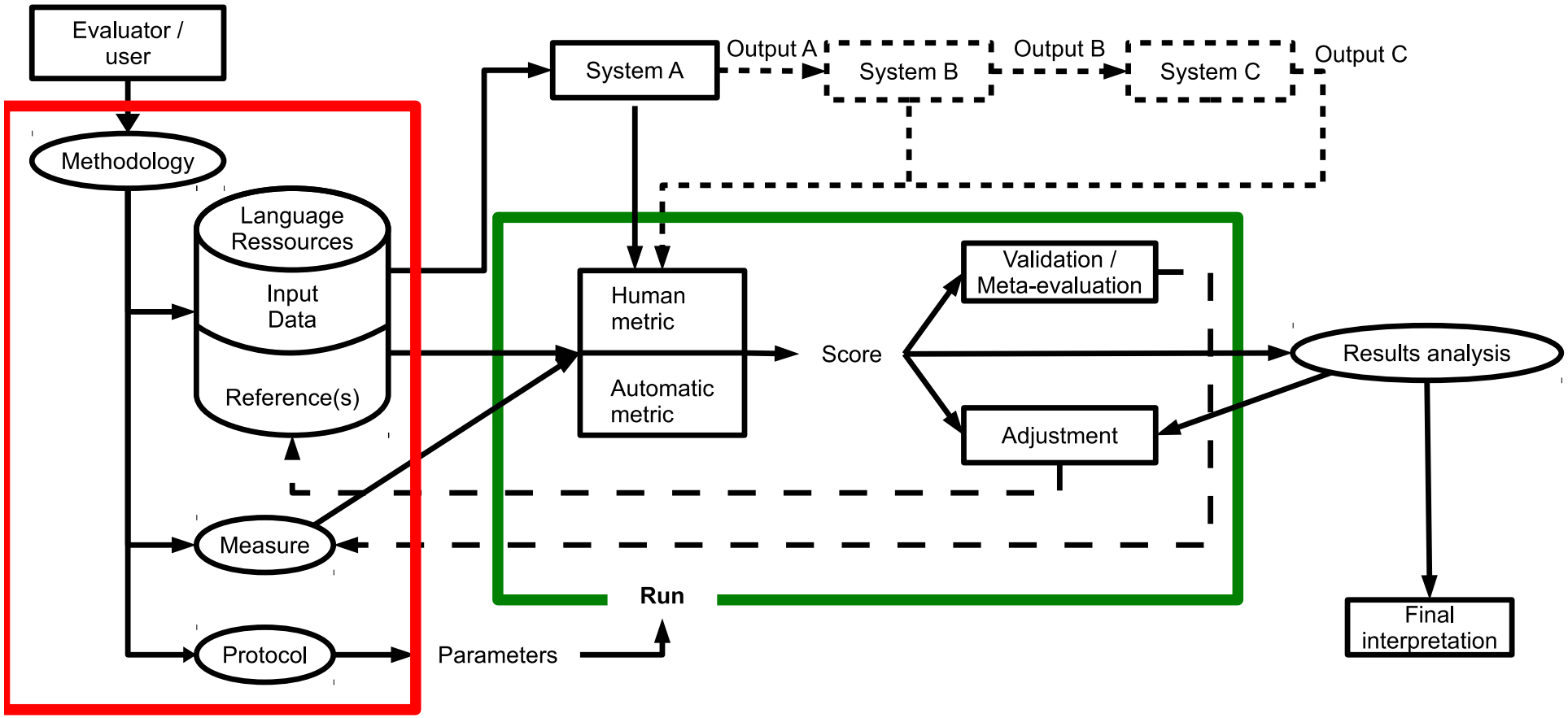
Evaluation Workflow



Evaluation Workflow



Methodology



Two evaluation cycles

CLARA

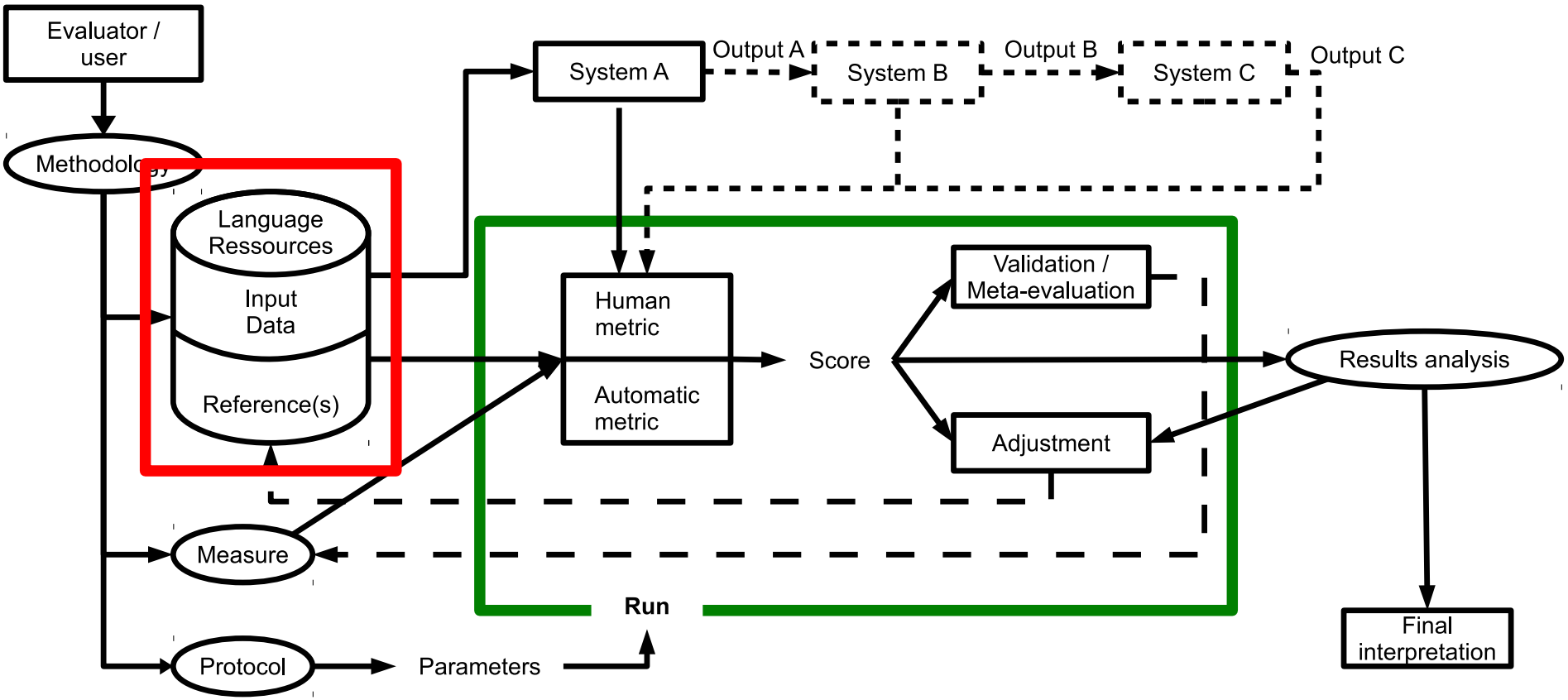
Preliminary cycle

- Define the protocol
- Set up the evaluation
- Set up tools (and platform...)

Practical cycle

- Apply the protocol, run the evaluation workflow
- Measure quality
- Analyse and interpret results

Language Resources



Language Resources

CLARA

Things to think about...

- Availability or state of a LR
- Reuse
- Intellectual Property Rights (IPR)
- Impact of the LRs on the evaluation results
- Reuse of tools/scripts to build new resources

Reuse of LRs regarding the technology

Technology	Monolingual				Multilingual			
	Lex.	Term.	Spch Corp.	Text Corp.	Lex.	Term.	Spch Corp.	Text Corp.
Spell checking	X							
Machine Trans.				X	X	X		X
Terminological Extr.		X		X				
Information Extr.	X	X			X			X
Automatic Summar.				X				
Document Index.	X				X			
Information Retr.		X		X		X		X
Speech Recog.			X				X	
Speech Synth.			X					

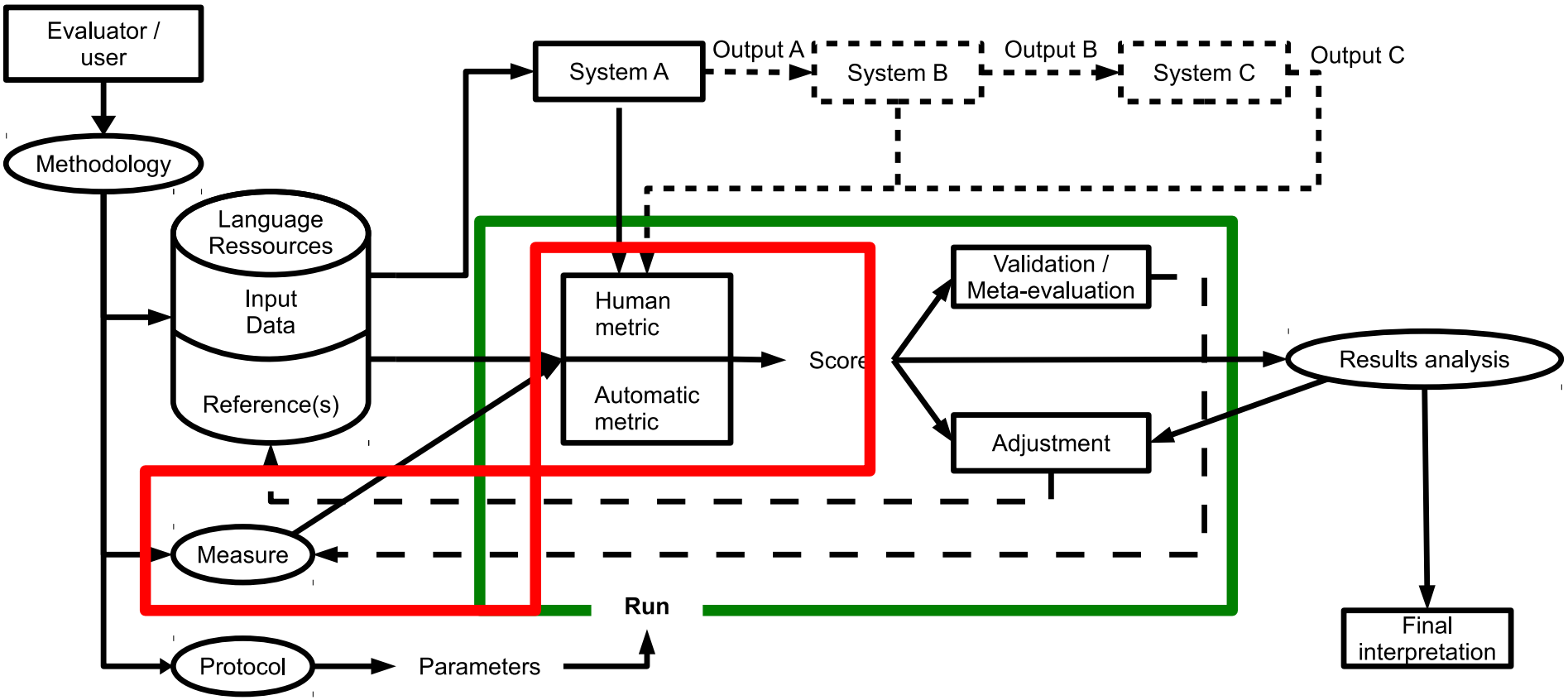
LR life cycle example

CLARA

French/English parallel corpus on medical domain

- EQueR/EVALDA project : question-answering evaluation
- CESART/EVALDA project : terminological extraction evaluation
- CESTA/EVALDA project : machine translation evaluation
- Could have been used for alignment evaluation...

Evaluation measures



Measure choice

CLARA

Two types

- Human
- (Semi-)automatic

Human part

- Always there
  - Human judgements
  - Reference
- Otherwise, systems could integrate the measure

Small number of measures

Technology	Human measures		automatic measures	
	n values	binary	distance	Precision / Recall F-measure
Automatic translation	X		X	X
Question-answering	X			
Synthesis	X			
Speech-to-speech translation	X			X
Terminologic extraction	X	X		X
Information retrieval		X		X
Speech recognition			X	
Alignement			X	
Automatic summary			X	
Parsing				X

Human measures

CLARA

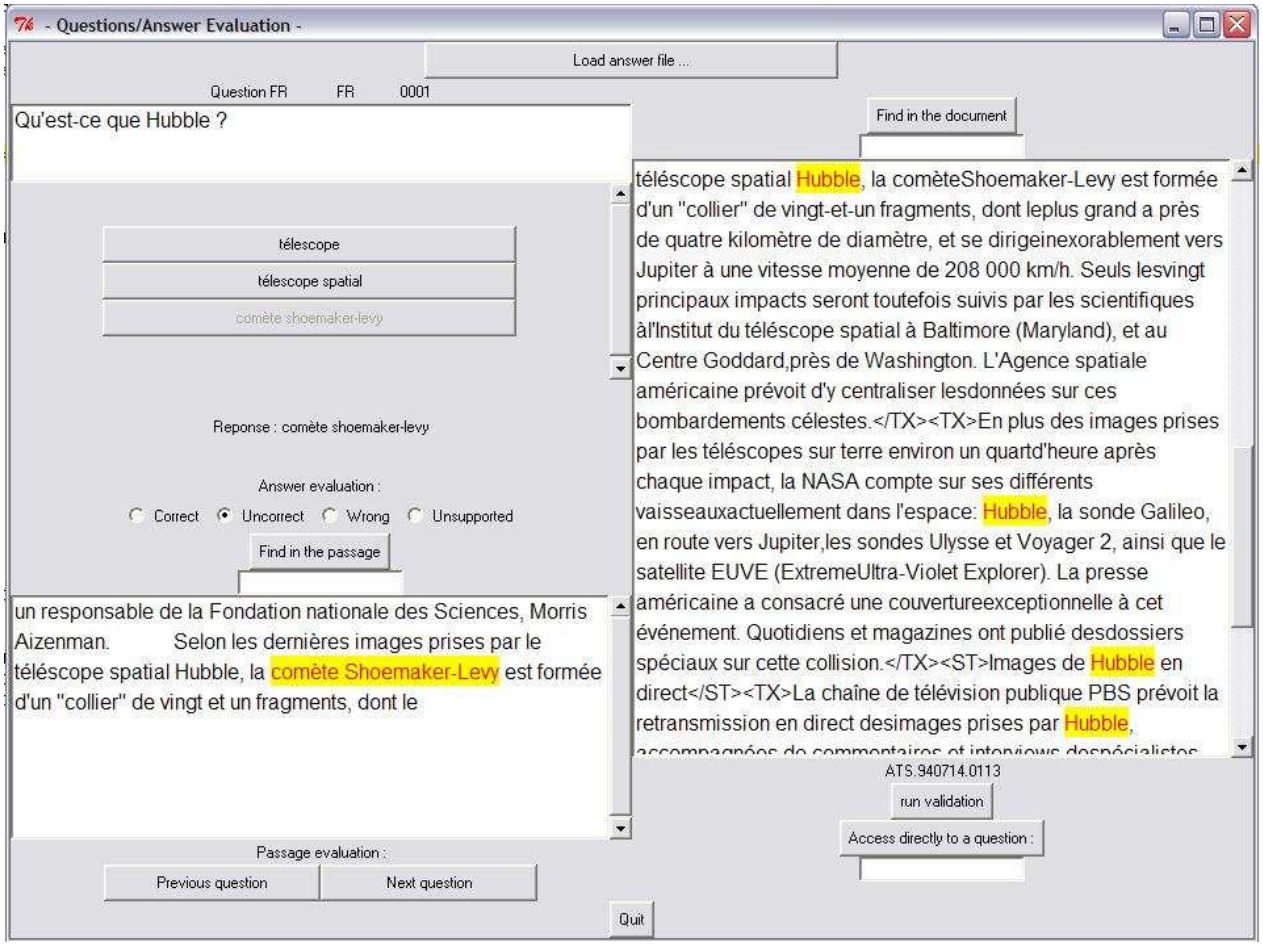
Characteristics

- Observations (judgements) made by human (judges)
- Variable subjectivity
- High cost ($, set up, delays)
- Good reliability iif judgements are well supervised

Needs

- Set up the evaluation
- Human judgements
- Interpretation and visualization of the results
- Evaluator follow-up

Interface : Question-answering



Interface : MT (set up)

Evaluations
Segments
Juges
Administration

Segments

ID	System	Document	Segment	Reference
1801	cesta-run2-hum-en	TEST-CESTA-DOC-12	18	no
1802	cesta-run2-hum-en	TEST-CESTA-DOC-12	180	no
1803	cesta-run2-hum-en	TEST-CESTA-DOC-12	181	no
1804	cesta-run2-hum-en	TEST-CESTA-DOC-12	183	no
1805	cesta-run2-hum-en	TEST-CESTA-DOC-12	184	no
1806	cesta-run2-hum-en	TEST-CESTA-DOC-12	185	no
1807	cesta-run2-hum-en	TEST-CESTA-DOC-12	186	no
1808	cesta-run2-hum-en	TEST-CESTA-DOC-12	187	no
1809	cesta-run2-hum-en	TEST-CESTA-DOC-12	188	no
1810	cesta-run2-hum-en	TEST-CESTA-DOC-12	189	no
1811	cesta-run2-hum-en	TEST-CESTA-DOC-12	19	no
1812	cesta-run2-hum-en	TEST-CESTA-DOC-12	190	no
1813	cesta-run2-hum-en	TEST-CESTA-DOC-12	191	no
1814	cesta-run2-hum-en	TEST-CESTA-DOC-12	192	no
1815	cesta-run2-hum-en	TEST-CESTA-DOC-12	193	no
1816	cesta-run2-hum-en	TEST-CESTA-DOC-12	194	no
1817	cesta-run2-hum-en	TEST-CESTA-DOC-12	195	no
1818	cesta-run2-hum-en	TEST-CESTA-DOC-12	196	no
1819	cesta-run2-hum-en	TEST-CESTA-DOC-12	197	no
1820	cesta-run2-hum-en	TEST-CESTA-DOC-12	198	no
1821	cesta-run2-hum-en	TEST-CESTA-DOC-12	199	no
1822	cesta-run2-hum-en	TEST-CESTA-DOC-12	2	no
1823	cesta-run2-hum-en	TEST-CESTA-DOC-12	20	no
1824	cesta-run2-hum-en	TEST-CESTA-DOC-12	200	no
1825	cesta-run2-hum-en	TEST-CESTA-DOC-12	201	no

Display from **1801** to **1825** (on **10153** evaluations)

<< Page 73 on 407 >>

delete | import

1822

delete

Les Directives d'évaluati
(version de 1996) seront
entrepris par un Groupe
conclu que ces directives
méthodologies toxicologi

Interface : MT (fluency)

Le texte est-il écrit en bon français ?

Et moyennant abonnement annuel 15 dollars par famille, ont pu bénéficier des résidents sont gratuitement des médicaments essentiels et le transport, le transfert à l´hôpital en cas d´urgence.

- Niveau 5 - Français parfait
- Niveau 4
- Niveau 3
- Niveau 2
- Niveau 1 - Français incompréhensible

segment suivant

Évaluations réalisées : 15 / 96

Interface : MT (adequacy)

A quel point le sens exprimé dans la traduction de référence est aussi exprimé dans la traduction cible ?

L´UNICEF est la force motrice qui contribuer à édifier un monde arriver des droits des enfants.

- Niveau 5 - Tous le sens
- Niveau 4
- Niveau 3
- Niveau 2
- Niveau 1 - Aucun sens

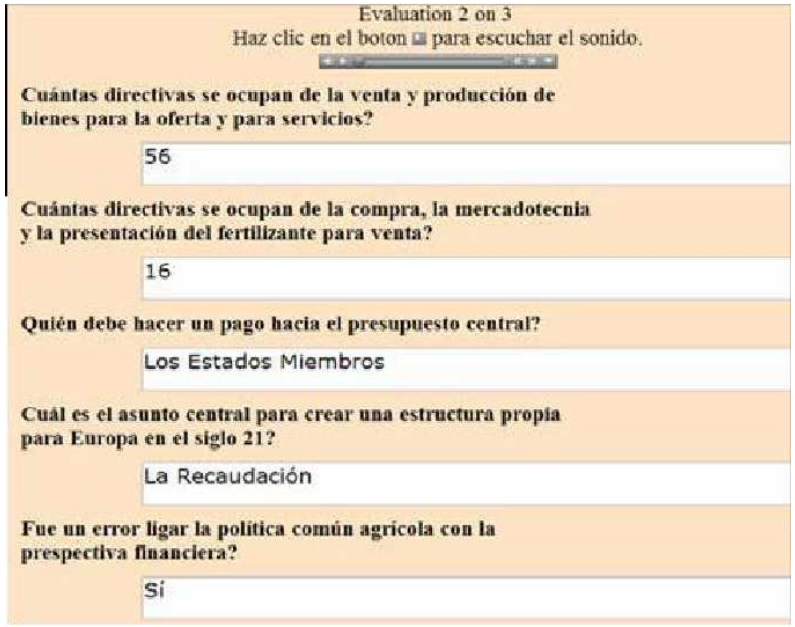
La traduction de référence est la suivante:

L´UNICEF est l´élément moteur qui aide à construire un monde où les droits de chaque enfant seront réalisés.

segment suivant

Évaluations réalisées :29 / 96

Interface : Speech-to-Speech translation



(Semi-)automatic measures

CLARA

Characteristics

- Comparison with one or several references
- Objectives : replace human judgements
  - when they are not "possible"
  - when they are too costly
- Advantages : execution speed, reproductibility, *objectivity*, *cost*, workflow integration

Integration and automation : various ways...

CLARA

Scripts

- Simple and fast implementation
- Task merging, reproductibility at wish

Integration and automation : various ways...

```
mylaptop$ perl 01_CHECK_SUBMISSIONS.PL
[...]
mylaptop$ perl 02_LIST_SUBMISSIONS.PL
[...]
mylaptop$ perl 03_EVALUATION.PL
[...]
mylaptop$ perl 04_BUILD_RESULTS_TABLES.PL
[...]
```

Integration and automation : various ways...

CLARA

Evaluation platforms

- Evaluation results tracking
- Genericity
- Easy evaluation access
- *Need some programming*
- Users do (most of) the job

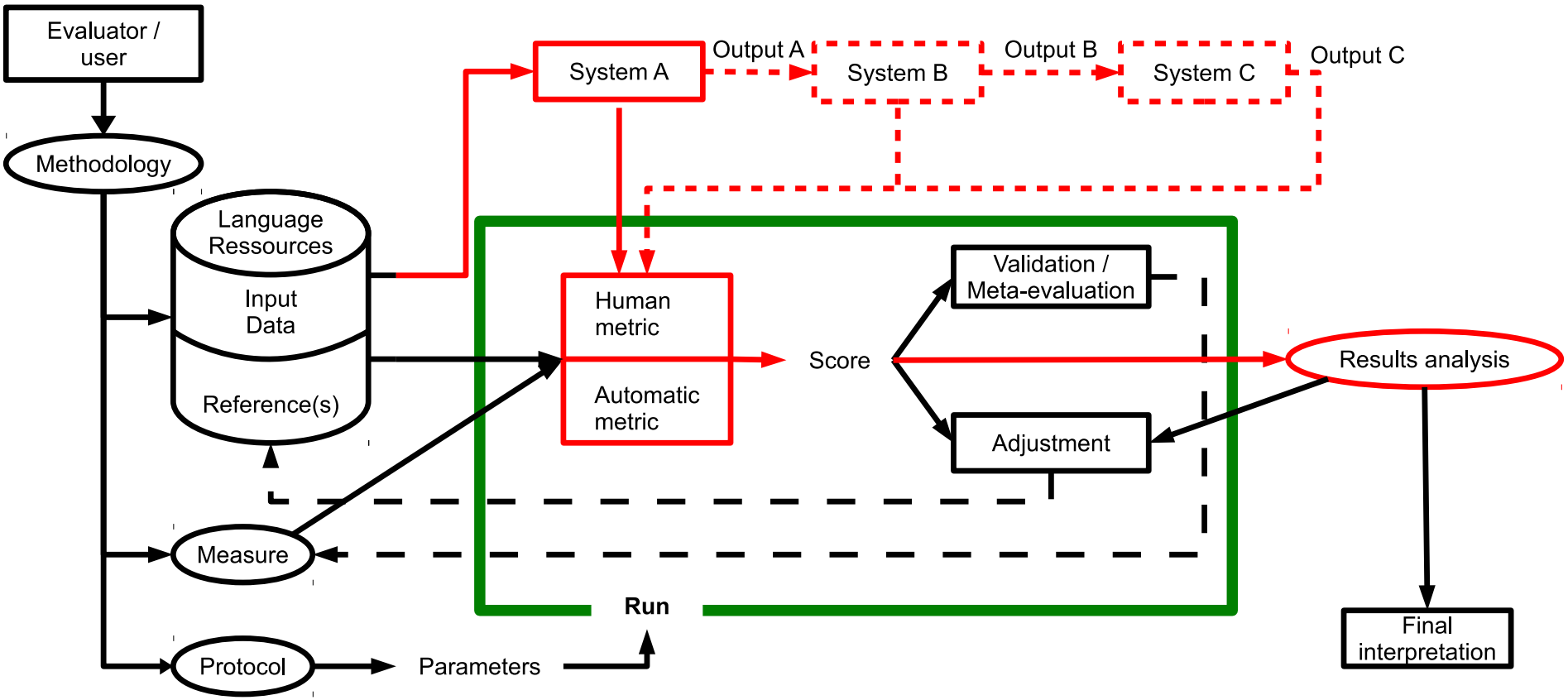
Integration and automation : various ways...

CLARA

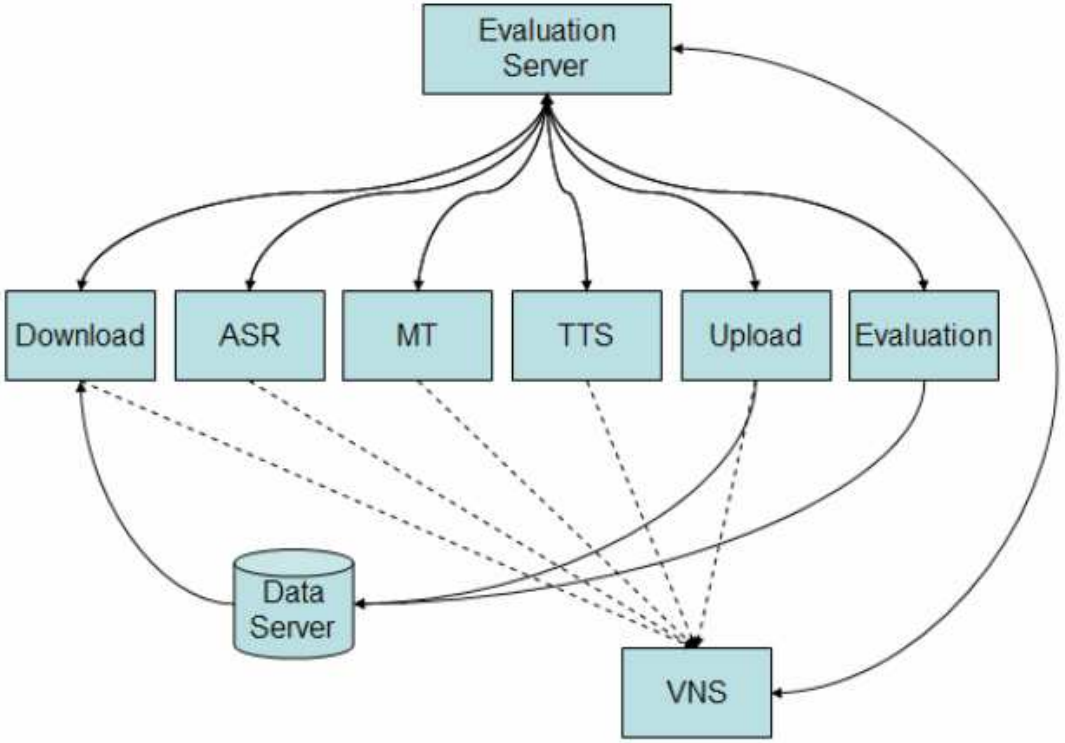
Example at Elda

- Speech-to-Speech translation
  - TC-STAR project
  - UIMA usage
  - 3 technical components, evaluation components

Speech-to-Speech evaluation - TC-STAR



Speech-to-Speech evaluation - TC-STAR



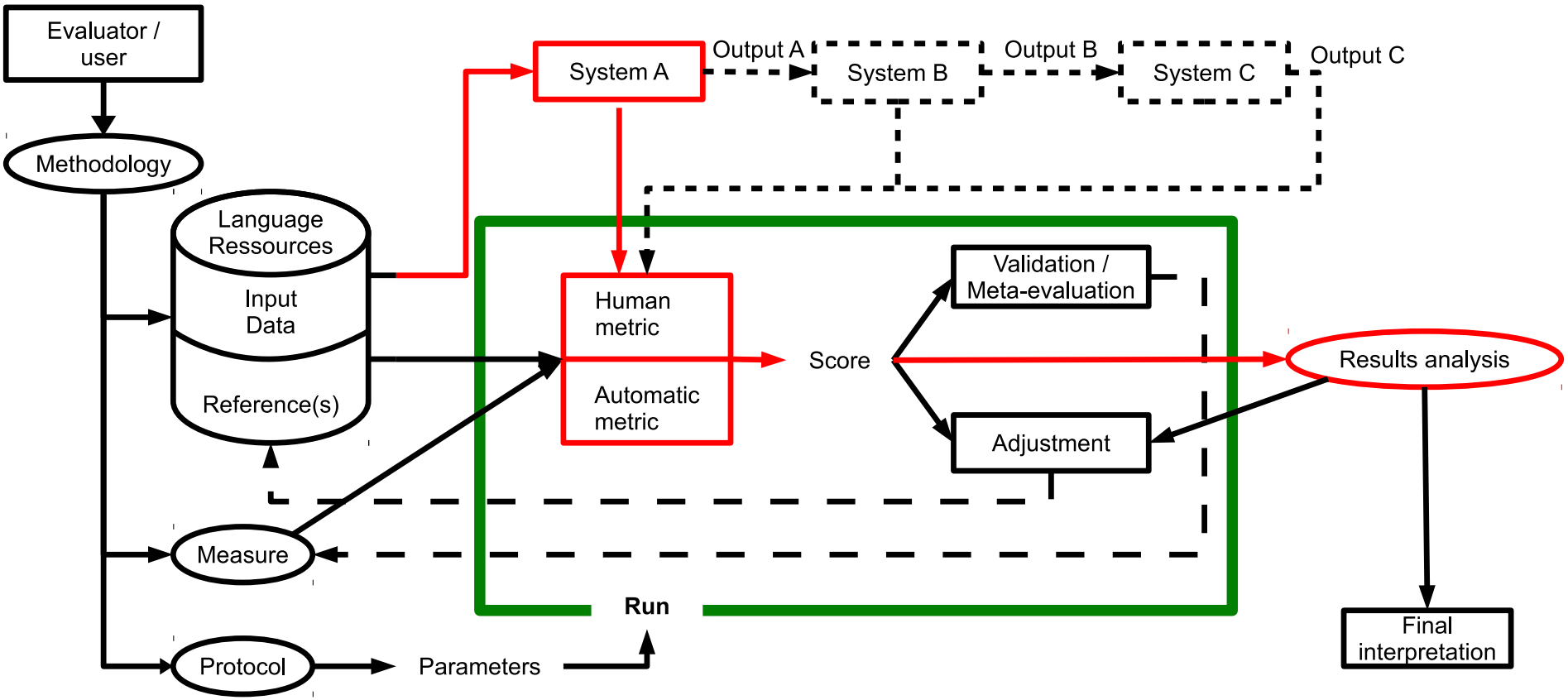
Integration and automation : various ways...

CLARA

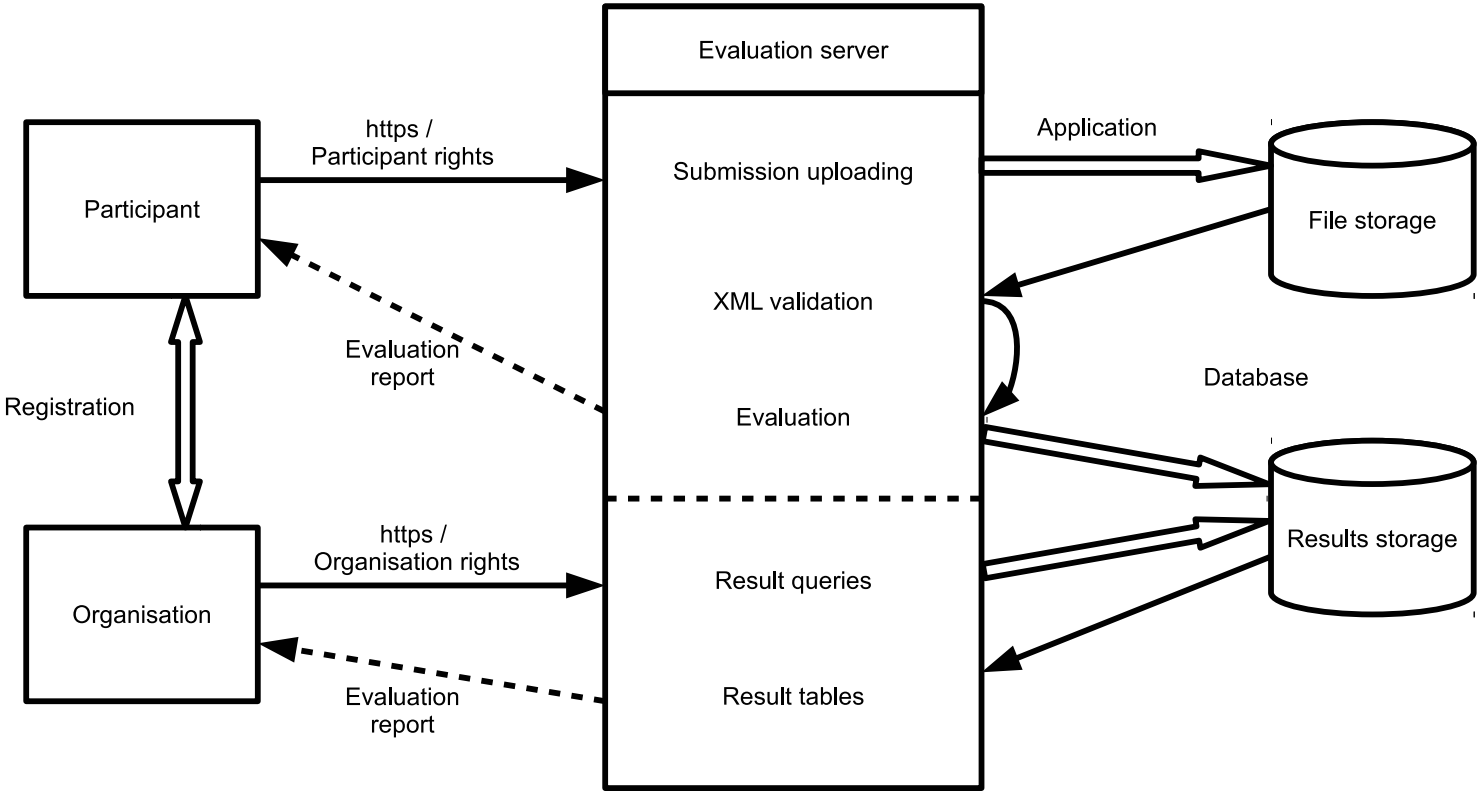
Example at Elda

- Parsing
  - PASSAGE project
  - Open access to the server during the system development cycle
  - Two evaluation campaigns run

Parsing evaluation within PASSAGE



PASSAGE server



PASSAGE : Results visualization

Campagne :campagne_passage

Utilisateur : admin_passage
Description de l'évaluation :

Numbre d'évaluations réalisées : 8

Date de l'évaluation : 2008-01-09

Heure de l'évaluation : 13:11:08

Parcourir

Précision moyenne pour tous les corpus, tous les constituants : 0.880327

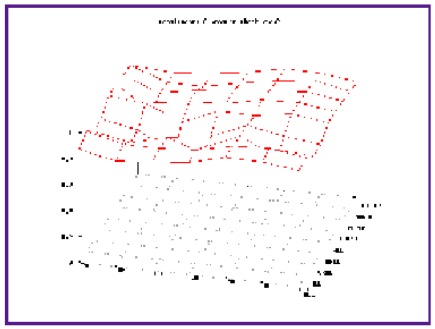
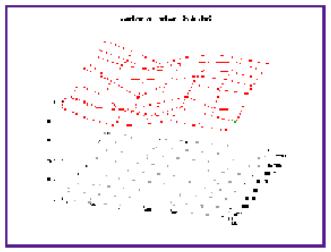
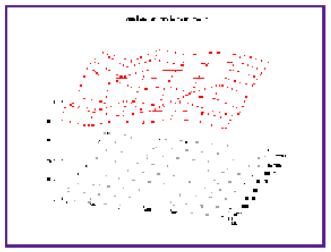
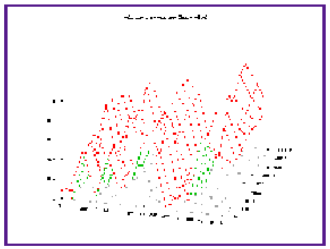
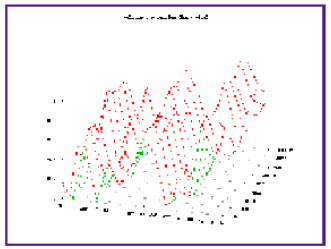
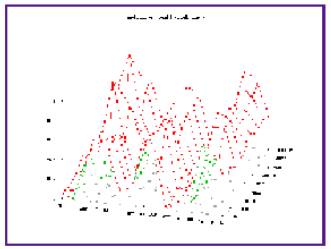
Rappel moyen pour tous les corpus, tous les constituants : 0.894776

F-mesure moyenne pour tous les corpus, tous les constituants : 0.887493

Précision moyenne pour tous les corpus, toutes les relations : 0.53164

Rappel moyen pour tous les corpus, toutes les relations : 0.369984

F-mesure moyenne pour tous les corpus, toutes les relations : 0.43632

```
CONSTITUENT EVAL SUBCORPUS_/tmp/EASYEVAL_Wed_Jan__9_13_09_18_CET_2008/ref_general_elda NV p=0.954248  r=0.875 f=0.912908
CONSTITUENT EVAL SUBCORPUS_/tmp/EASYEVAL_Wed_Jan__9_13_09_18_CET_2008/ref_general_elda GA p=0.766355  r=0.836735 f=0.8
CONSTITUENT EVAL SUBCORPUS_/tmp/EASYEVAL_Wed_Jan__9_13_09_18_CET_2008/ref_general_elda GR p=0.904762  r=0.926829 f=0.915663
CONSTITUENT EVAL SUBCORPUS_/tmp/EASYEVAL_Wed_Jan__9_13_09_18_CET_2008/ref_general_elda GP p=0.934959  r=0.98 f=0.95695
CONSTITUENT EVAL SUBCORPUS_/tmp/EASYEVAL_Wed_Jan__9_13_09_18_CET_2008/ref_general_elda PV p=1  r=0.888889 f=0.941176
EVAL SUBCORPUS_/tmp/EASYEVAL_Wed_Jan__9_13_09_18_CET_2008/ref_general_elda ALL_RELATIONS p=0.553191  r=0.404255 f=0.467139
```

PASSAGE : follow-up for the evaluator

CLARA

Campagne :campagne_passage

Nom d'utilisateur : admin_passage
Nombre d'évaluations réalisés : 8

Résumé de vos évaluations primaires (la plus récente est probablement la plus juste) :

#Eval Primaire	Date	Heure	Constituants			Relations			Détails
			F-mesure	Précision	Rappel	F-mesure	Précision	Rappel	
0	2007-12-22	00:46:44	0	0	0	0.039553	0.650602	0.0203965	Détails
1	2008-01-03	16:07:57	0	0	0	0.039553	0.650602	0.0203965	Détails

Résumé de vos évaluations de développement:

Evaluation id	Description	Date	Heure	Constituants			Relations			Détails	Pri
				F-mesure	Précision	Rappel	F-mesure	Précision	Rappel		
1		2007-11-29	18:23:34	0	0	0	0.506128	0.536541	0.478979	Détails	
16		2007-12-10	14:15:46	0	0	0	0.506128	0.536541	0.478979	Détails	
19		2007-12-10	14:25:53	0	0	0	0	0	0	Détails	
20		2007-12-10	14:38:02	0	0	0	0.506128	0.536541	0.478979	Détails	
26		2007-12-13	10:17:49	0.0303321	0.552605	0.015594	0.0122559	0.322148	0.00624676	Détails	
27		2007-12-13	10:54:51	0	0	0	0.506128	0.536541	0.478979	Détails	
28		2007-12-13	10:56:25	0	0	0	0.0268061	0.595265	0.0137118	Détails	
2		2008-01-09	13:11:08	0.887493	0.880327	0.894776	0.43632	0.53164	0.369984	Détails	

Integration and automation : various ways...

CLARA

Web services / automatic workflow

- Not that hard to implement
- When available, easy use
- Flexible
- Allows a free access to the tools

Integration and automation : various ways...

CLARA

Example

- PANACEA project
- Tools available through web services (registry.elda.org)
- Build workflows from the available web services (myexperiment.elda.org)

PANACEA : build a service with an ACD file

Using a Tomcat server (http ://tomcat.apache.org)

```
appl: BLEU_Evaluation [
  documentation: "BLEU Evaluation within MEDAR"
  groups: "MEDAR"
  nonemboss: "Y"
  executable: "perl"
]

string: script  [
  standard: "Y"
  parameter: "Y"
  default: "/home/olivier/[...]/mteval-v11b.pl"
  comment: "display false"
  comment: defaults
]

boolean: bool_env  [
  additional: "Y"
  information: "case sensitive evaluation"
  qualifier: "c"
  default: false
]
```

```
infile: reference_file  [
  standard: "Y"
  qualifier: "r"
  comment: "data direct"
]

infile: source_file  [
  standard: "Y"
  qualifier: "s"
  comment: "data direct"
]

infile: target_file  [
  standard: "Y"
  qualifier: "t"
  comment: "data direct"
]

outfile: output  [
  additional: "Y"
  default: "stdout"
]
```

PANACEA : run a workflow

CLARA

Using Taverna (http ://www.taverna.org.uk)

- (demo)

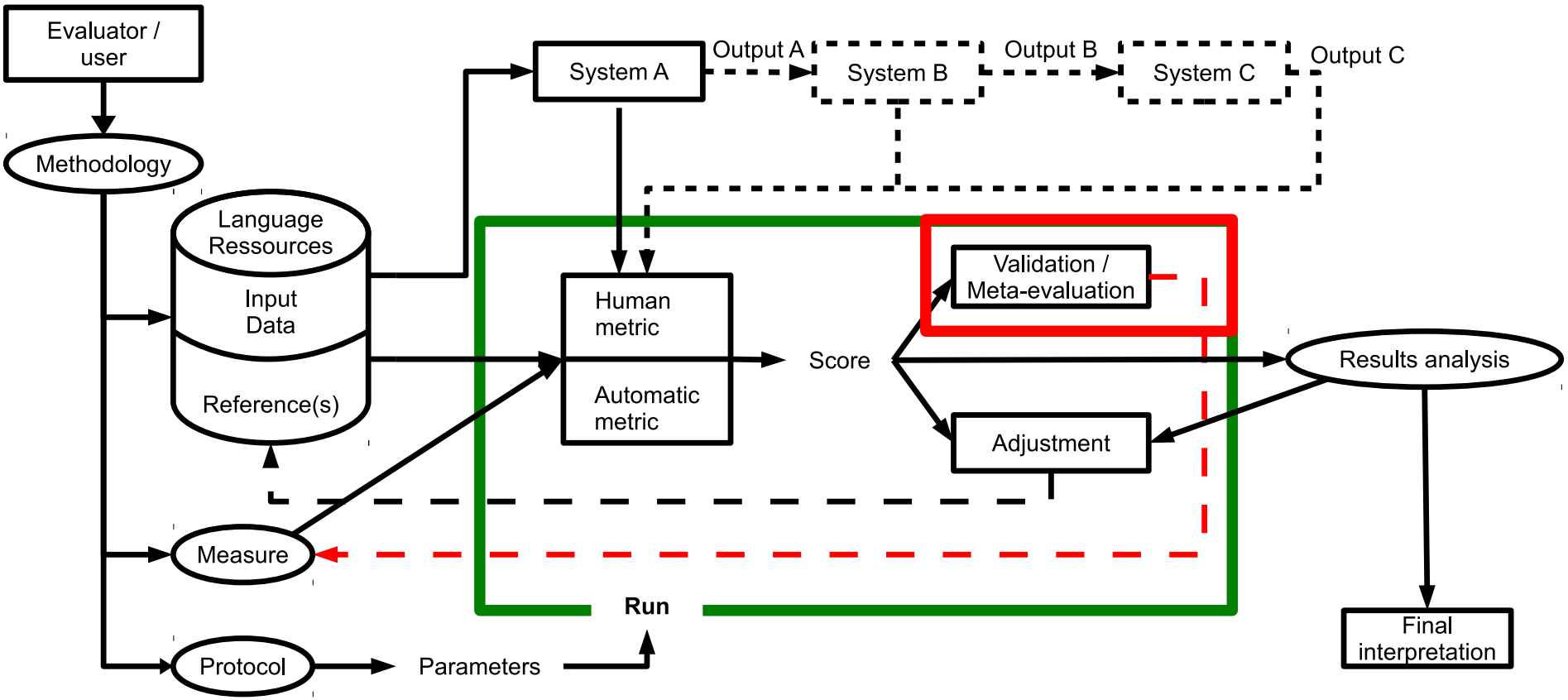
Integration and automation : various ways...

CLARA

What shall I implement ? It depends on :

- the size of the evaluation (versions of one system, a whole evaluation campaign, etc.)
- the usage (by the evaluator vs by the system developers...)
- the repetitivity of the evaluation (3-year project, a development evaluation once a week, etc.)
- my knowledge
- my available time

Validation and meta-evaluation



Validation and meta-evaluation

CLARA

Automation

- Human judgements : validation
- Automatic metrics : meta-evaluation
- (Don't forget to meta-evaluate to check metrics !)

Validation of human measures

CLARA

- Measure agreement (Kappa coefficient, inter-judge and intra-judge agreements)
- Allow to interpret results
- Identify diverging judges

Meta-evaluation of metrics

CLARA

Fiability : relevance of scores

- Comparison with another *reference* measure
- Correlation coefficient (Pearson, Spearman, Kendall)

Robustness : production of similar scores for data of similar quality

- Data samples (*bootstrapping*)
- Difference with the samples' mean

Conclusions

CLARA

How to be lazy ?

- Maximum reuse of the existing
- Do not reinvent the wheel
- Avoid duplicated tasks/tools (DRY)

How to be productive

- Build as many as possible generic things
- Use fast methodologies and tools
- For new metrics, be creative !
- Do not forget : quality of the results is the final objective