

ICIS Data Validation Tool

Maria Corina D. Habito
IRRI-CRIL

Agenda

- Objective of the tool
- Background
- What's New?
- What's Next?

Objective of the tool

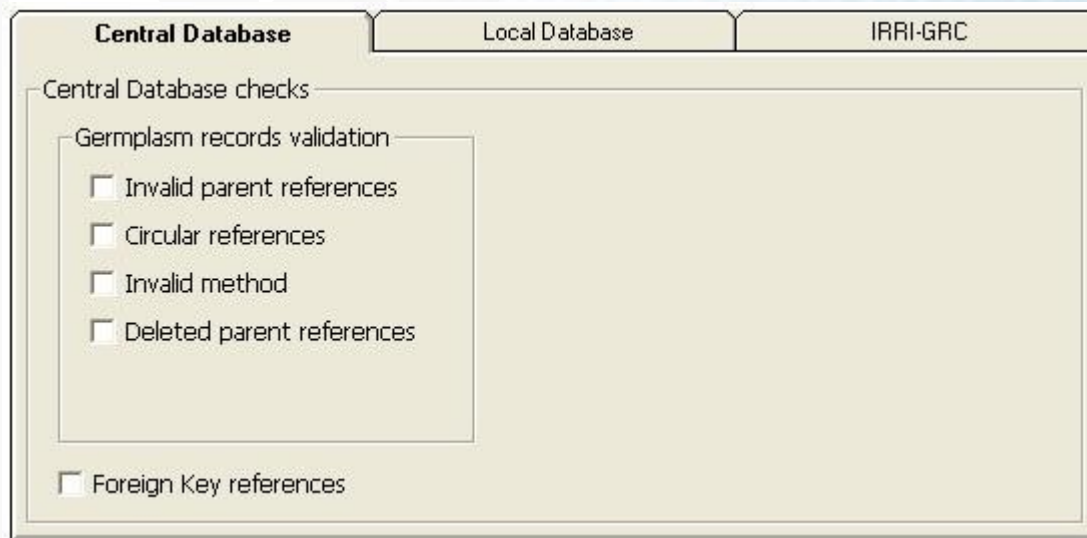
- To search ICIS for data errors that might render it meaningless.

Background

- First introduced at 2006 ICIS Developers' Workshop
- Originally developed using Visual Basic

Background

- No. of central database checks: 11 (5 checkboxes)

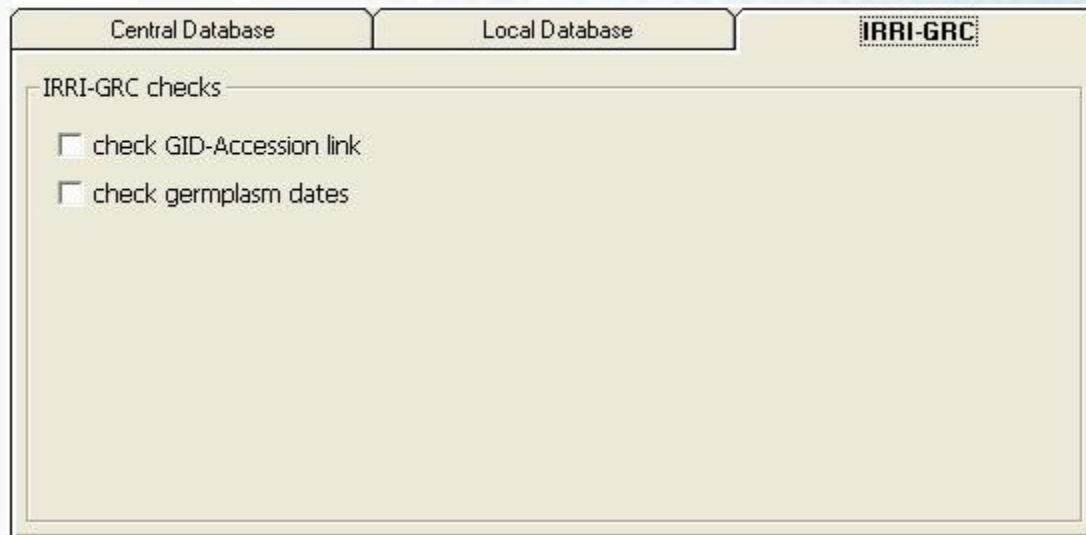


The screenshot shows a software interface with three tabs: "Central Database", "Local Database", and "IRRI-GRC". The "Central Database" tab is active. Inside this tab, there is a section titled "Central Database checks" which contains a sub-section "Germplasm records validation" and a "Foreign Key references" checkbox. The "Germplasm records validation" sub-section contains four checkboxes: "Invalid parent references", "Circular references", "Invalid method", and "Deleted parent references". All checkboxes are currently unchecked.

Database Type	Check Name	Checked
Central Database	Invalid parent references	<input type="checkbox"/>
	Circular references	<input type="checkbox"/>
	Invalid method	<input type="checkbox"/>
	Deleted parent references	<input type="checkbox"/>
	Foreign Key references	<input type="checkbox"/>

Background

- 2 checks specific to IRRI-GRC data (2 checkboxes)

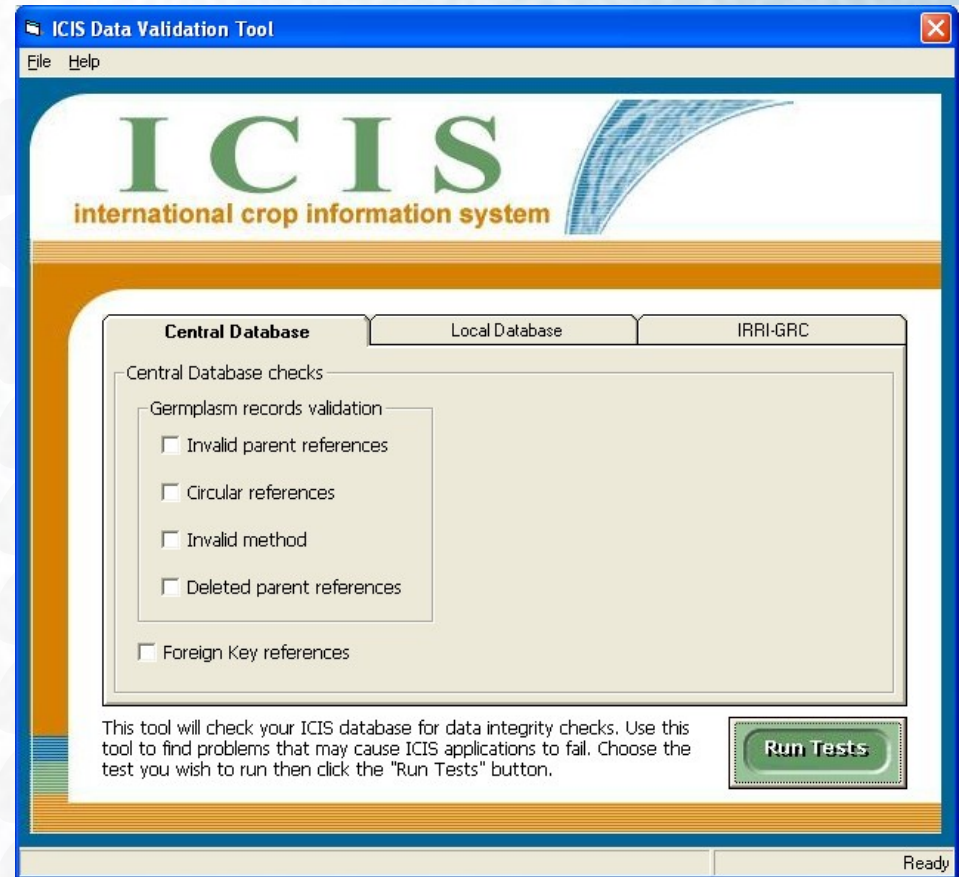


The screenshot shows a software window with three tabs: "Central Database", "Local Database", and "IRRI-GRC". The "IRRI-GRC" tab is selected. Below the tabs, there is a section titled "IRRI-GRC checks" containing two checkboxes:

- check GID-Accession link
- check germplasm dates

Background

- Original user interface:

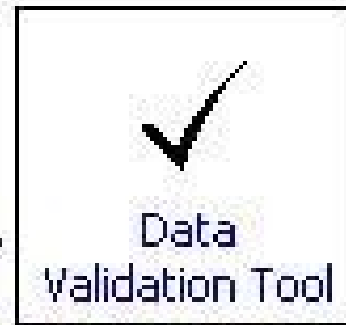
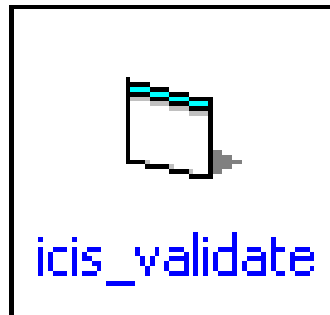


Agenda

- Objective of the tool
- Background
- What's New?
- What's Next?

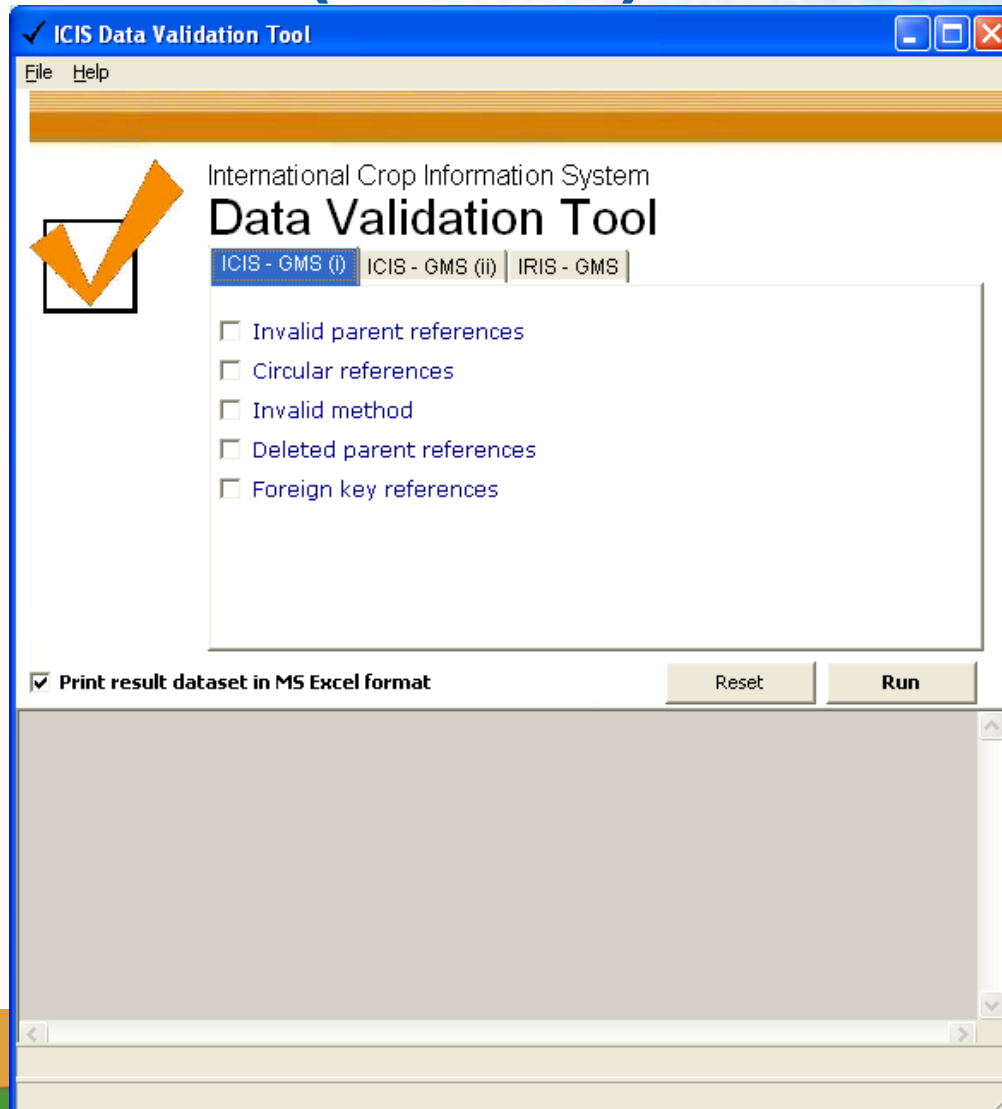
What's New?

- Original Visual Basic prototype translated to Delphi (commenced Jan 2007)
- Application icon



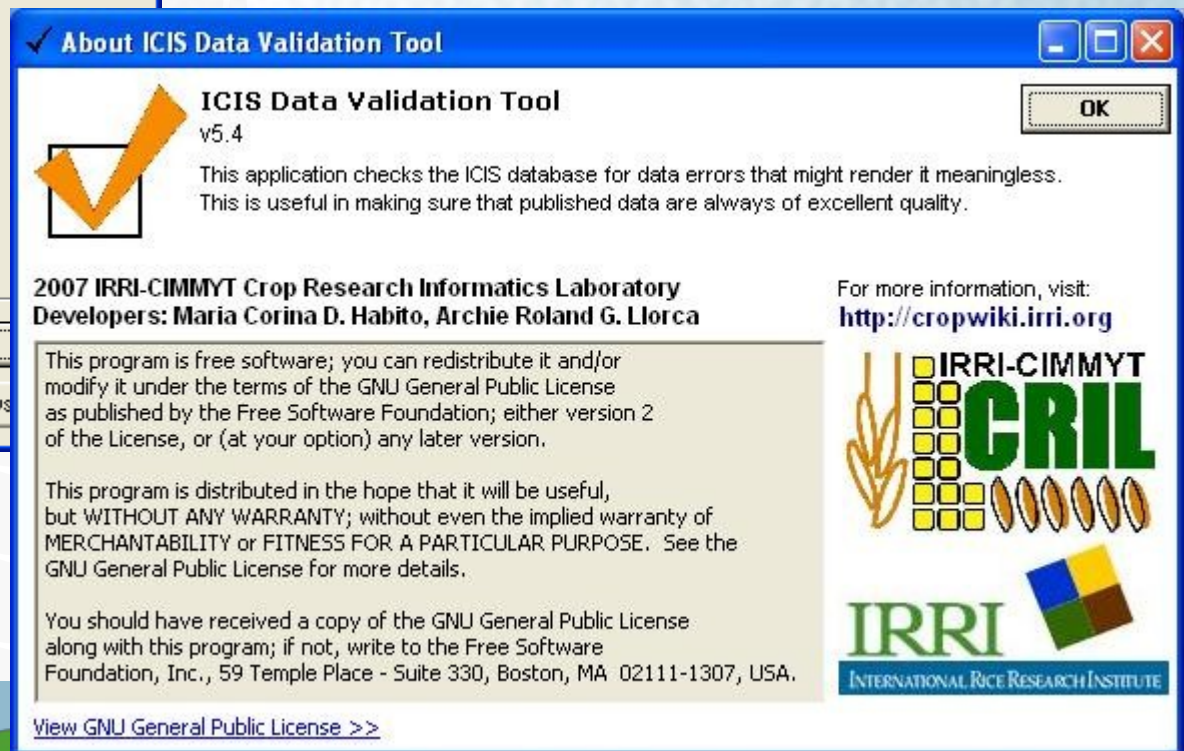
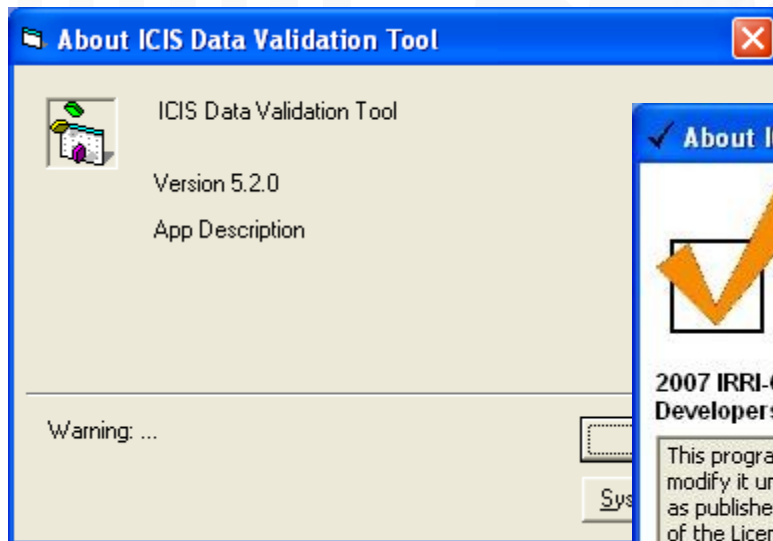
What's New? (cont'd)

- New look:



What's New?

- Option to output results in Excel
- Improved “About” form



What's New?

- 19 additional data checks (ICIS:6, IRIS:13)

Progenitor germplasm dates Progenitor ID1 unknown, Progenitor ID2 known Name inheritance from GPID2: check NDATE and NLOCN

New data checks

Progenitor germplasm dates [3 checks]

(The GDate of a GID must not predate GDATE of any of its progenitors)

New Check #1: GDATE earlier than GDATE of GPID1 (Error-0012)

New Check #2: GDATE earlier than GDATE of GPID2 (Error-0013)

New Check #3: GDATE earlier than GDATE of MGID (Error-0014)

New data checks

Progenitor ID1 unknown, Progenitor ID2 known
(A GID can't have GPID2>0 and GPID1=0)

New Check #4: Progenitor ID1 unknown, Progenitor ID2 known (Error-0015)

Progenitor germplasm dates Progenitor ID1 unknown, Progenitor ID2 known Name inheritance from GPID2: check NDATE and NLOCN

New data checks

Name inheritance from GPID2: check NDATE and NLOCN [2 checks]

New Check #5: NDATE not inherited (Error-0016)

New Check #6: NLOCN not inherited (Error-0017)

- Preferred Names
- Preferred IDs
- Location (GLOCN) of germplasm
- Method - name type combinations
- Name type occurrence
- Name type RELNM (Release Name)

New data checks

Preferred Names [2 checks]

New Check #7: More than one preferred English name (Error-0018)

New Check #8: With invalid name type as preferred name. Names eligible are: CRSNM, RELNM, DRVNM, CVNAM, ELITE (Error-0019)

Preferred Names Preferred IDs Location (GLOCN) or germplasm Method - name type combinations Name type occurrence Name type RELNM (Release Name)

New data checks

Preferred IDs [3 checks]

New Check #9: More than one preferred ID (Error-0020)

New Check #10: With invalid name type as preferred ID. Eligible to be preferred ID: DRVNM, COLNO, ACCNO (if present), GACC (if present), ITEST (if present), CIATGB (if present) (Error-0021)

New Check #11: Two accessions must not share the same name of the same type if one or both is a preferred ID. (Error-0022)

ICIS - GMS (i) | ICIS - GMS (ii) | IRIS - GMS

 Preferred Names Preferred IDs Location (GLOCN) of germplasm Method - name type combinations Name type occurrence Name type RELNM (Release Name)

New data checks

Location (GLOCN) of germplasm

New Check #12: With GLOCN different from GLOCN of GPID2 [for Method <> IMPORT] (Error-0023)

- Preferred Names
- Preferred IDs
- Location (CLOCN) of germplasm
- Method - name type combinations**
- Name type occurrence
- Name type RELNM (Release Name)

New data checks

Method - name type combinations [3 checks]

METHOD TYPE	VALID NAME TYPES
GEN (Generative)	CRSNM, UNCRS, UNRES
DER (Derivative)	RELNM, DRVNM, CVNAM, CVABR, NTEST, LNAME, ADVNM, ACVNM, AABBR, OLDMUT1, OLDMUT2, ELITE, UNRES
MAN (Management)	ACCNO, RELNM, CVNAM, CVABR, COLNO, FACCN, ITEST, NTEST, LNAME, TACC, ADVNM, ACVNM, ELITE, GACC, DACCN, LCNAM, CIATGB

New Check #13: Invalid name type for method GEN (Error-0024)

New Check #14: Invalid name type for method DER (Error-0025)

New Check #15: Invalid name type for method MAN (Error-0026)

- Preferred Names
- Preferred IDs
- Location (GLOCN) of germplasm
- Method - name type combinations
- Name type occurrence
- Name type RELNM (Release Name)

New data checks

Name type occurrence

New Check #16: With name types occurring more than once (Error-0027)

(ACCNO, CRSNM, UNCRS, COLNO, ITEST, GACC, CIATGB, RELNM, DRVNM, CVNAM)

New Check #17: Cross-names & Line-names can't occur together (Error-0028)

New Check #18: Release names & Collector's Nos. can't occur together (Error-0029)

ICIS - GMS (i) | ICIS - GMS (ii) | IRIS - GMS

- Preferred Names
- Preferred IDs
- Location (GLOCN) of germplasm
- Method - name type combinations
- Name type occurrence
- Name type RELNM (Release Name)

New data checks

Name type occurrence

New Check #19: Germplasm sharing a release name with another germplasm in the SAME country (Error-0030)

Agenda

- Objective of the tool
- Background
- What's New?
- What's Next?

What's Next?

- Implementation of remaining data checks suggested by Dr. Hamilton (for IRIS)

CropForge Feature Requests Search [Advanced search](#)

Home My Page

Summary Forums Tracker Lists Docs

[Feature Requests: Browse](#) | [Download .csv](#) | [Submit New](#) | [Reporting](#) | [Monitor](#) | [Admin](#)

[#160] Data validation checks

Monitor (?)

Submitted By: Ruraidh Sackville Hamilton (rhamilton)

Data Type: (?) Feature Requests

Group: None

Assigned To: (?) Nobody (Admin)

State: (?) Open

Summary: (?) Data validation checks

Date Submitted: 2006-05-21 12:00

Category: General ICIS

Priority: (?) 3

[Submit](#) [Delete this item](#)

Detailed description

Following the demo of the data validation checker in CIMMYT, I'd like to register the following if not already in:

- General validation checks for all ICIS implementations
 - GID:GDate must not predate GDate of any of its progenitors (n.b. need to check all and MGID, to deal properly with missing GDATES e.g. GID:GDate = YYYYMMDD, GID:GPID2:GDate > YYYYMMDD
 - If GID:NAME:NVal=GID:GPID2:NAME:NVal, then GID:NAME:Ndate=GID:GPID2:NAME:Ndate and GID:NAME:NLocN=GID:GPID2:NAME:NLocN
 - A GID can't have GPID2>0 if GPID1=0
 - Only 1 name of a GID can have NSTAT=1
 - If GERMPISM could have a new field to identify the location of the germplasm holder distinguish GLOCN = the location of germplasm production) then one could validate the GLOCN. In this case, GID:GLOCN should equal GID:GPID2:germplasm holder.

Talk:Data Validation Tool - ICISWiki

article discussion edit history move watch

Talk:Data Validation Tool

Contents [hide]

- 1 Validation checks for all ICIS implementations
- 2 Validation checks for IRIS
- 3 Desirable data validation not yet possible
- 4 Validation checks for IRGC accessions

Validation checks for all ICIS implementations

- The GDate of a GID must not predate GDate of any of its progenitors. (Beware missing links in the chain of dates! If t and MGID, then it will not detect the following error: non-zero GPID1, GPID2 or MGID has GDATE=0 but their GPID1 GPID1, GPID2 or MGID has GDate=0, iterate to check their GPID1, GPID2 and MGID)
- If a GID inherits a name from its source (GPID2) then that name record must also inherit the NDATE and NLOCN
- A GID can't have GPID2>0 and GPID1=0

Validation checks for IRIS

- Exactly one name (neither more nor less) of a GID must have NSTAT=1 (preferred English name)
- Names eligible to be the preferred name are: CRSNM, RELNM, DRVNM, CVNAM, ELITE
- No more than one name of a GID can have a "preferred ID" status (NSTAT =8)
- The following name types, if present, must be preferred ID: ACCNO, GACC, ITEST, CIATGB
- The following names types are eligible to be preferred ID: DRVNM, COLNO
- The preferred ID must be unique for the given name type – that is, two accessions must not share the same name of
- For all methn#62, GLOCN of a GID = GLOCN of its GPID2
- Only certain combinations of name type and germplasm creation method are acceptable, namely
 - Method type GEN: valid name types are: CRSNM, UNCRS, UNRES
 - Method type DER: valid name types are: RELNM, DRVNM, CVNAM, CVABR, NTEST, LNAME, ADVNM, ACVNM
 - Method type MAN: valid name types are: ACCNO, RELNM, CVNAM, CVABR, COLNO, FACCN, ITEST, NTEST, LCNAM, CIATGB
- Name type ACCNO, CIATGB and ITEST must
 - have METHN=62
 - GLOCN = location of holding organization

What's Next?

- Add prefix of “DATA” to Error codes (e.g. DATAError-0001)
- Implementation of data checks for other ICIS implementations
- Allow checking of data in different database backends