线上故障管理规范V1.0

线上故障管理规范

第一章　总则

- 为规范外研在线业务系统生产环境故障管理，针对线上故障定级，规范处理流程等，特制订此规范。
- 故障定义：业务系统发布生产环境并验收通过后，一些事件发生导致线上业务无法正常开展，用户体验受到破坏，此为故障。
- 制订本规范期望达到的目标：

1. 恢复业务运行
2. 减少类似故障
3. 建立知识库，利于后续维护
4. 提高在用户之前发现故障的比例

第二章 故障发现

- 故障发现途径：

1. 系统监控告警。监控项指标异常，触发告警。
2. 主动巡检。技术人员查看生产环境日志，或者例行检查监控项时，看到了异常，进而发现了故障。
3. 用户发现。用户使用时发现了问题，反馈给公司客服等人员，客服上报问题。

第三章 故障等级

- 业务系统分级：须产品部门提交相关文档，确定业务系统是否属于公司主业务，描述系统的主、次流程，系统的主、次功能点，并动态更新。据此及影响范围认定故障等级及处理要求、恢复时限。

故障等级	严重程度	描述	处理要求			业务恢复时限
			运维处理	项目组处理	公司级处理	
S0	灾难问题，须立即处理	公司全域产品崩溃； 影响所有用户。	/	问题发生后，项目组负责人负责问题定位和处理，如果1小时内无法定位问题，需要上报到公司级进行处理	指定合适的人员参与直至问题解决	<1小时
S1	严重问题，尽快处理	1个主业务或多个非主业务产品线系统崩溃、主流程阻塞等问题； 影响多校用户	/	问题发生后，项目组负责人负责问题定位和处理，如果2小时内无法定位问题，需要上报到公司级进行处理	指定合适的人员参与直至问题解决	<2小时
S2	重要问题	1个主业务产品线1个次要流程阻塞； 1个非主业务产品线主流程阻塞； 影响多校用户	问题发生后，运维作为负责人负责问题定位和处理，如果2小时内无法定位问题，需要上报到项目组负责人	项目组组织人员进行问题定位和修复处理，如果4小时内依然无法定位问题，需要上报到公司级进行处理	指定合适的人员参与直至问题解决	<4小时
S3	一般问题	1个主业务产品线1个主要功能异常； 1个非主业务产品线1个次要流程阻塞； 影响1校用户	问题发生后，运维作为负责人负责问题定位和处理，如果4小时内无法定位问题，需要上报到项目组负责人	项目组组织人员参与直至问题解决	/	<8小时
S4	轻微问题	1个主业务产品线1个次要功能异常； 1个非主业务产品线1个主要功能异常； 影响少量用户	问题发生后，运维作为负责人负责问题定位和处理,直至问题解决	/	/	<12小时

第四章 故障通知

- 通知对象：技术、产品、客服及支持同事，对应负责人，中心/事业部领导，公司领导等。
- 故障处理负责人或指定协调人须在故障认定、故障恢复、故障解决时通过公司工作群等方式通知相应同事，通知内容至少包括故障描述、影响范围、当时进展等。
- 各故障等级须通知对象：

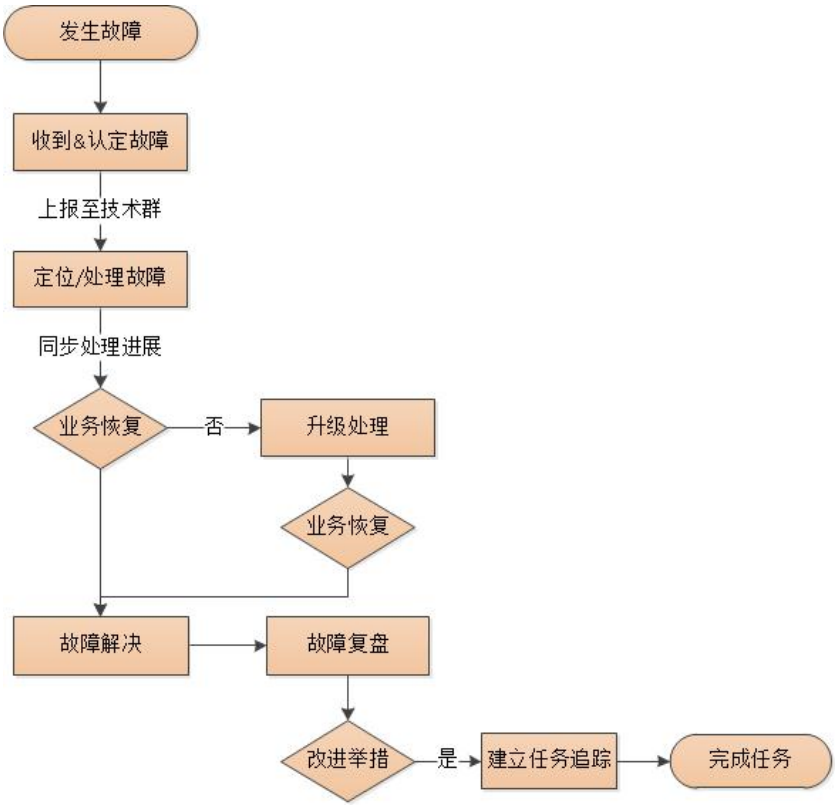
通知对象	技术			产品			业务支持中心					市场		公司负责人
	对应研发同事	技术负责人	公司CTO	对应产品同事	产品负责人	事业部负责人	值班客服	对应支持同事	客服负责人	支持负责人	中心负责人	对应市场同事	市场负责人	
S0	√	√	√	√	√	√	√	√	√	√	√	√	√	√
S1	√	√	√	√	√	√	√	√	√	√	√	√	√	
S2	√	√		√	√		√	√	√	√		√	√	
S3	√			√			√	√				√		
S4	√						√							

第五章 故障处理

- 故障处理原则：发生故障时须尽快恢复业务，避免或减少故障带来的损失。

1. 一切以恢复业务为最高优先级。第一时间恢复业务，而不是彻底解决问题。
2. 当前负责人不能短时间内解决，则必须升级处理。
3. 处理过程在不影响用户体验的前提下，保留现场。

- 故障处理流程图：



- 处理过程中的反馈机制：故障处理负责人或指定协调人每隔 10～15 分钟在技术群或临时组建的故障处理群做一次反馈，反馈当前处理进展以及下一步行动。中途有需要立即执行的操作，须事先通报操作对业务系统的影响是什么，由负责人决策后再执行。
- 故障复盘：故障解决后，Confluence填写《线上服务故障统计》表；一周内组织相关人员对故障进行复盘（S0和S1故障必须复盘），建议采用COE分析方法，形成文档存入公司知识库。同时对故障等级进行认定，以及团队成员责任的归属。
- 改进举措：针对当前故障要做哪些改进措施，应对类似问题如何预防等。需要给出可实施的方案以及时间计划。
- 完善系统运维文档：对系统运维文档查漏补缺进行完善。相关人都可以依据运维文档，针对常见故障能紧急处理恢复业务。运维文档常规内容如下：

1. 管理平台使用文档，包括服务启动/停止/重启等操作
2. 服务程序部署、回滚版本操作流程等
3. 服务程序部署目录、日志目录列表
4. 系统资源扩容操作
5. 常见故障列表及恢复手册

第六章 附则

本规范适用于外研在线技术共享中心与各事业部技术部。

本规范的解释权归外研在线技术共享中心。

附件

附件1 线上服务故障统计

序号	影响的产品	故障发现来源	故障开始时间	响应时间	恢复时间	故障等级	影响业务和用户描述	故障原因分析	解决方案	后续预防/改进措施

附件2 故障COE分析方法

项目	描述
标识	<业务标识>-<等级S0/S1/S2>-<日期>-<顺序号>
状态	草稿/执行中/结项
问题描述	
影响业务和客户范围	
原因分析（鱼骨图/因果分析图）4-6层原因	
经验总结：规则，经验，知识	
改进措施	
处理过程（指标，图表，代码）	